

How not to make friends

The United States has rejected an attempt to give teeth to the Biological Weapons Convention, complaining that the protocol is too weak. If so, this is largely because of US demands. No wonder America's allies feel betrayed.

Last week's announcement that the United States will not support a proposed verification regime for the international treaty banning biological weapons did not come as a surprise. It was, however, profoundly disappointing to the negotiators who have been working on the document over the past six years.

Those negotiations began after it became clear that the 1972 Biological Weapons Convention (BWC) was being flouted by some of its signatories. In 1989, a defector from the Soviet Union revealed that it was supporting a bioweapons programme ten times larger even than Western intelligence agencies had feared. A few years later, United Nations inspectors discovered facilities for mass-producing anthrax in Iraq.

For months, US negotiators have been making it clear that they are unhappy with the proposed verification scheme that has emerged (see *Nature* **411**, 223 & **412**, 365; 2001). One argument has been that its provisions are too weak. But this is disingenuous. The United States has repeatedly opposed the attempts of its allies in the European Union to strengthen the inspection provisions of the protocol. The current draft, for instance, includes loopholes allowing some facilities that might conceivably be used for bioweapons production to be excluded from the inspection regime.

Other countries had conceded to these demands in the hope of salvaging the agreement. They realized that the protocol was never likely to detect every violation of the BWC. Better, they thought, to agree a relatively relaxed inspection regime, and strengthen it later on, than to fall at the first hurdle. This is why last week's announcement has left the United States' allies with a bitter taste in their mouths.

It is also difficult to see how the United States is serving its own national interest. True, the US pharmaceutical industry had voiced concerns that inspections could put proprietary information at risk. But it also indicated that it was ready to thrash out a workable solution. This should not have been that hard, given that European companies were prepared to go along with stricter inspection regimes outlined in earlier drafts of the protocol. The US chemical industry has also lived since 1997 with the more rigorously policed Chemical Weapons Convention.

The other reason to scrap the protocol is to protect from prying eyes the United States' research into defences against bioweapons. This has been a long-standing concern of the chief US negotiator, Donald Mahley. Although Mahley defied President Bill Clinton's pro-protocol policy, he was never replaced. Now he has President George W. Bush's backing.

US negotiators have made tentative suggestions about the sort of verification they would like to see, involving such notions as getting the World Health Organization to monitor for suspicious outbreaks of disease. Such alternatives are viewed as unacceptably vague by most signatories to the BWC.

The Bush administration appears to believe that it can rely on its biodefences and let the BWC go to the wall. Similar logic underlies the decision to press ahead with the National Missile Defense system. This suggests a naive faith in technological solutions and a reckless disregard for the value of arms control. Jeopardizing multilateral arms-control treaties makes the entire world a more dangerous place in which to live. Scientists who believe in arms control should redouble their efforts to press this case. ■

Time for orbiting lab to find true purpose

The International Space Station's current plight provides another chance to subject its research portfolio to rigorous review.

Watching the apparent ease with which NASA astronauts have been assembling the International Space Station in orbit, one would scarcely suspect the chaos the programme is experiencing down on Earth. NASA is requesting up to an additional \$5 billion over the next five years. The US Congress, furious with the space agency's failure to control the project's costs, seems unlikely to oblige. If so, science stands to be the loser, with lab facilities and astronaut-hours for research likely to be cut (see page 465).

With so much invested and already built, and with so many countries committed to the project, cancellation is not a viable option. But so badly have NASA's managers bungled their cost estimates that no one can say with certainty what the final tab will be.

Meanwhile, scientists wait nervously to see what will remain when the dust settles. Congress is clearly on record saying that research is a primary purpose of the station, and no one wants it to be merely an orbiting hotel. So what, after all this time and trouble, is its scientific *raison d'être*?

Until the original plans were scaled back by Congress, President

Ronald Reagan's expansive 1980s vision for the space station, then dubbed 'Freedom', included everything from gravitational biology, through Earth observation, to commercial 'factories' making super-size protein crystals that would lead to breakthroughs in drug design.

That last rationale survives in more modest form, but has been challenged repeatedly, most recently last year by a US National Academy of Sciences expert panel that found little merit in past space experiments in protein crystal growth. The academy called on NASA either to demonstrate the potential for significant scientific advance or to drop such experiments from the space station.

This kind of hard-headed thinking is exactly what is needed as scientists sit down to help NASA decide what to throw out should science have to be scaled back. The current crisis at least provides an opportunity to look again at the station's research portfolio, and to concentrate on those areas that are widely agreed to show the most promise. If all that happens is a self-interested scramble among lobbyists representing the different research communities involved, it will be an unfortunate end to a sorry tale. ■





Bushwhacked
Herbal remedy lands
Johns Hopkins in
hot water
p466



Infection agents
Disease detectives
throw light on
BSE's origins
p467



Silent witness
Missile-defence critic
in secrecy row with
Pentagon officials
p468



Missing millions
Embezzlement
rocks Weizmann
Institute
p469

Tough decisions loom as funding crisis hits space-station research



Tony Reichhardt, Washington

Following massive cost overruns, NASA is considering a drastic cut-back on research equipment aboard the International Space Station. The agency's proposals have angered scientists, particularly those who feel that their plans to study how biological systems adapt in space are under threat.

Kathie Olsen, who heads the agency's Office of Biological and Physical Research, says she will hold a workshop "as soon as possible" to enlist advice from the scientific community for working out a reduced 10-year plan for space-station science.

Olsen's office has responded to a White House order to bring budgets into line (see *Nature* **410**, 399; 2001) by proposing a 40% reduction in funding for laboratory equipment. This would entail scrapping key pieces of lab equipment due to be installed after 2004, including an animal holding cage and facilities for cell culture and plant studies. The cuts would decimate gravitational biology studies on the station, say researchers in the field.

Meanwhile, another space-station partner, Japan, is having difficulty building a 2.5-metre centrifuge. This is a critical piece of hardware that would allow scientists to simulate different levels of gravity on test plants and animals. Japan's project managers have pressed NASA either to scale back its techni-

cal requirements for the centrifuge or to come up with more money.

But the worst cut-back, say scientists, would be if NASA abandoned work on a space-station rescue vehicle. This would effectively reduce the onboard crew from seven people to three. With only three astronauts on board, says Kenneth Baldwin of the University of California at Irvine, who heads NASA's advisory committee for biological research, "you can't do science, I don't care what anyone says". He estimates that these measures would eliminate 75–85% of the basic research planned on the station.

A three-person crew would also mean that European and Japanese researchers would have little chance of working in orbit, leaving the project's international partners anxious about their role in a scaled-down station. Science investigations for the facility are being selected jointly by the partners, who had also expected to share NASA lab equipment. Olsen has been talking to science managers at the European Space Agency about flying European-built space-station equipment earlier than planned to help compensate for the proposed NASA cuts.

In deciding which US equipment should be scrapped, NASA science managers gave priority to research on astronaut health and biotechnology over basic research in gravitational biology and materials science. That

Going overboard: cuts could mean that NASA's space-station 'lifeboat' would be abandoned.

NASA/AP

Nuclear physicists red-faced over elementary mistake

David Adam, London

Physicists at the Lawrence Berkeley National Laboratory in California have been forced to withdraw claims that they had discovered two new 'superheavy' elements. The embarrassing U-turn follows several failed attempts to reproduce the original results in further experiments at Berkeley as well as at laboratories in Germany and Japan.

The researchers' claims to have observed two new elements with 118 and 116 protons were made in a 1999 paper published in the journal *Physical Review Letters* (83,

1104–1107; 1999). The heaviest artificial element previously generated, which was created in Russia, had 114 protons. The lead author of the paper, Victor Ninov, was quoted at the time as saying "we jumped over a sea of instability onto an island of stability that theories have been predicting since the 1970s".

But the researchers have now been left with wet feet and red faces, and have retracted the paper. Re-analysis of their original data using different software revealed little sign of the much-vaunted element 118. "There are

many lessons here and the lab will extract all the value it can from this event," says Charles Shank, lab director at Berkeley. "The path forward is to learn from the mistakes."

Other physicists say that they were always sceptical about the claims, pointing out that there was little genuine proof in the original paper. "You could believe or not believe, but physics is not a matter of belief," says Gottfried Münzenberg, leader of the department of nuclear structure and nuclear chemistry at the German national laboratory for heavy-ion research in Darmstadt. ■



rankled with some space biologists, who point to studies questioning the value of past biotechnology experiments in space (see *Nature* **404**, 114; 2000).

If NASA must choose between disciplines to make financial ends meet, it should first establish "what is the mission of the station", says Milburn Jessup, a surgeon at the University of Texas Health Science Center in San Antonio, who chairs NASA's station utilization subcommittee. Before the workshop is held to outline research priorities, the agency should decide which of two broad thrusts is the more important — a long-term presence in space, including human missions to the Moon and Mars, or research on improving products and processes on Earth.

Such a choice may be necessary if Congress decides not to give NASA the extra \$4 billion–\$5 billion the agency says it needs over the next five years to complete the station — this is on top of the \$8 billion already budgeted. Early signs are that lawmakers are so exasperated with the agency's poor control over costs that they may withhold the money as a disciplinary measure. In a recent report accompanying the NASA spending bill for 2002, the Senate appropriation subcommittee that oversees the agency's budget said in unusually blunt terms: "The Committee has lost confidence in the [station] programme's ability to responsibly manage the budget."

But Baldwin concludes that more money for the rescue vehicle and other hardware is "the only solution we see" that would ensure that the station is an efficient multidisciplinary laboratory. Martin Fettman, who chairs NASA's oversight panel for some of the lab equipment marked for cancellation, says his committee "has come close" to resigning in protest over the proposed cuts. Although scrapping equipment such as the animal holding facility is "unacceptable", he says, scientists might be willing to support stretching out the timeline for installing such equipment on the station. But until NASA comes back with an amended plan, he says, "we're out in the vapour, waiting".

Johns Hopkins embroiled in fresh misconduct allegations

K. S. Jayaraman, New Delhi

A researcher from Johns Hopkins University in Baltimore is among those accused of unethical conduct in a clinical trial carried out in India. The scandal has emerged barely a week after clinical trials at Johns Hopkins University were temporarily suspended following the death of a volunteer in a drug trial (see *Nature* **412**, 363; 2001).

The high court in the state of Kerala has accepted a petition claiming that an investigational drug was injected into cancer patients at the Regional Cancer Centre (RCC) in Thiruvananthapuram, the state capital, without clearance from the Drug Controller General of India, and without approval from the institutional ethical committee.

The chief minister of Kerala, Arakkal Parambil Kurien Antony, last week ordered an investigation into the clinical trial.

The drug in question is M4N, a derivative of the natural product nordihydroguaiaretic acid (NGDA), which has been used as a herbal remedy but is known to cause liver and kidney damage. Krishnan Nair, RCC director and principal investigator of the clinical trial, confirmed that the drug was injected into tumours of 27 patients awaiting surgery between November 1999 and April 2000. Clearance from the Drugs Controller was first received in 2001. Ganga Devi, convener of the ethical committee, said that approval was given in November 1999 for topical drug administration but not for injection.

But Nair told *Nature* that he had "verbal" permission from the Drug Controller General and written consent from patients. He said he thought that the drug was non-toxic, on the basis of studies in mice conducted at Johns Hopkins, which had also indicated that it was effective against virus-induced cancers in mice.

Ru Chih C. Huang, a professor of biology at Johns Hopkins, supplied the drug, which his team had developed as part of an anti-HIV research programme. He also initiated and funded the Indian trial. Although details of the agreement have been kept secret, sources said that the RCC undertook this work for a fee of US\$750,000. Huang did not respond to requests from *Nature* for an interview.

Nair said that his US collaborators would be flying to India next week to respond to the allegations.

Narayanan Bhattathiri, associate professor of radiotherapy at RCC, blew the whistle on the trial. Although he is also head of the clinical radiobiology section, he says: "Cur-

ously we were not informed about the trial."

Bhattathiri informed the Human Rights Committee, a body established by the Indian parliament, that "what was done in the name of clinical research is very surprising". Each patient was given three injections into one area of their tumour. The tumour was surgically removed after three or four days for microscopic examination of the injected region. "It is clear that the injection was done not with the aim of producing any tumour regression," says Bhattathiri, "but to see how long the chemical stays in the injected spot and to evaluate its effect at a microscopic level."

Three patients whose tumours could not be removed were given radiation therapy, despite the fact that "no one knows the interaction between radiation and this drug," Bhattathiri says. "Such conduct in clinical trials will make Kerala and India an animal house."

But Nair said that he is happy with the outcome of the trials, claiming that the tumour in one patient disappeared three days after the injection. In a press statement on 28 February he claimed joint credit with Johns Hopkins University in announcing that the drug had been effective in treating certain virus-induced cancers.

But Bhattathiri says that the only role the RCC played in the development of the drug was to provide unsuspecting patients, who believed they were being treated with the latest therapy from the United States.



Trials of M4N, a drug derived from the creosote bush (*Larrea tridentata*), have raised concerns.

W. W. LOCKWOOD/CORBIS

Review blames BSE outbreak on calf feed

David Adam, London

Researchers studying the early days of the bovine spongiform encephalopathy (BSE) epidemic may have solved the mystery of why the disease developed in Britain, and why it emerged in the 1980s.

A review of the origins of BSE released on 19 July reveals that British farmers changed the diets of young calves to routinely include meat and bone meal (MBM) in the 1970s.

BSE is known to have been spread by contaminated MBM. The mixture has been regularly fed to older cattle around the world since the 1950s; feeding it to calves seems to have occurred in Britain, but not in the rest of Europe or the United States.

The review modifies the conclusions of last year's BSE inquiry, which suggested that the disease may have been caused by a freak genetic mutation.

The new review's finding suggests that young calves may be more susceptible than older cows to BSE. Epidemiologists studying the disease already suspected this, with some saying the risk to calves is 30 times higher. The new review recommends that possible age-related susceptibility should be tested.

Last year's inquiry, chaired by the judge Lord Phillips (see *Nature* 408, 3–6; 2000), failed to pick up the fact that feeding MBM to calves was largely restricted to Britain. "There was no attempt to conceal anything, it's just that the right questions never came up," says animal-feed consultant Brian Cooke, formerly of feed supplier Dalgety Agriculture, who presented evidence to the inquiry.

As a result, the finding has surprised many researchers. "I had not been aware that this was a practice introduced in the UK but had not been introduced elsewhere," says Peter Smith, an epidemiologist at the London School of Hygiene and Tropical Medicine and acting chairman of the government's Spongiform Encephalopathy Advisory Committee (SEAC).

Smith says that age-related susceptibility to BSE "must be a tenable hypothesis". He adds that parallels can be drawn to the emergence of the human equivalent of BSE, variant Creutzfeldt–Jakob disease (vCJD), which seems to affect mainly young people.

Smith does not think that this is because younger people are likely to have eaten more cheap pies and hamburgers — which are more likely to have contained potentially infectious brain and spinal cord. "I find it hard to get to grips with the idea that the age distribution in people is purely down to exposure," he says.

Why BSE should emerge in Britain alone when cows across the world had been eating MBM for decades puzzles researchers. Some say the disease originated from cows eating MBM prepared from sheep with a related



Feeding time: scrapie-infected meat and bone meal may have been fed to young calves.

encephalopathy called scrapie. Others argue that scrapie is so widespread it is inconceivable that such a jump would only happen in one place at one time, and so favour the 'freak genetic mutation' hypothesis.

This was the explanation favoured by the Phillips inquiry, which ruled out scrapie as the source. Many researchers felt this conclusion went beyond the available evidence. The new review was commissioned by the government to clear up this point, and concludes that scrapie cannot be excluded.

BSE is different from known strains of scrapie. But it might have been transformed in passing from sheep to cattle, or have been a rare strain that escaped researchers' notice.

"If you consider the properties of scrapie and the limited studies of the scrapie agent, one cannot rule out the unmodified scrapie agent in the way that Phillips did," says Gabriel Horn, professor of zoology at the University of Cambridge, who led the review committee.

One piece of evidence that implicates scrapie is that the only other country known to have introduced home-produced MBM to calf feed in the 1970s was Australia, which is free of scrapie — and of BSE.

But others are not convinced. "It would surprise me if MBM in calf feed was the sole reason why the BSE epidemic started in the UK and didn't appear anywhere else," says Malcolm Ferguson-Smith, a veterinary pathologist at Cambridge University and a co-author of the Phillips report. ■

Pressure grows over US blood ban

Sally Goodman, Paris

Blood-bank managers on both sides of the Atlantic are growing alarmed at the impact of proposed US restrictions on the importation of European blood.

The US Food and Drug Administration (FDA) and the American Red Cross are planning the restrictions as a precaution against the possible transmission through the blood supply of variant Creutzfeldt–Jakob Disease (vCJD), the human form of mad cow disease.

But the public organizations and charities that run the blood banks hope that researchers will quickly establish reliable diagnostic tests for vCJD in blood, so that the restrictions can be lifted.

One victim of the restrictions will be an agreement called Euroblood, under which blood products are exchanged between the United States and some European countries, currently the Netherlands, Germany and Switzerland.

Officials in New York city, which relies on blood imported from Europe for up to a quarter of its total supply, face an acute shortage as the restrictions come into effect. Robert Jones, president of the New York Blood Center, warns that the ban on

European blood would cause a medical crisis in the city. "We need to look at the relative risks here and balance these with the known risks of reduced blood supply," he says.

European officials say the restrictions are unnecessary and will undermine public confidence in the blood supply.

But reports that the restrictions will damage the European pharmaceutical industry by reducing supplies of blood plasma obtained from the United States under Euroblood are untrue, industry officials say, as the industry now imports plasma privately.

The FDA plans to implement the ban next year as part of broader guidelines to protect the US blood supply from vCJD that were proposed by an advisory committee last month (see *Nature* 412, 7; 2001).

The American Red Cross, which collects half of the US blood supply, is planning to exclude donors who have spent over six months in Europe since 1980, or three months in Britain. "We believe this will be an interim action, hopefully soon eclipsed by scientific information," says Bernadine Healy, president of the organization. "When we have a blood test it may be appropriate to modify this geographic deferral." ■

Hopes of biotech interest spur Latvian population genetics

Alison Abbott, Munich

Latvia has become the latest small European country to seek to gather genetic information on its population, in an attempt to entice the interest of international researchers and biotechnology firms.

A law being considered there would set guidelines for collecting medical and genomic information from the population, as well as data on its lifestyle. The guidelines are based on those used in neighbouring Estonia, where a similar programme is under way. Iceland has already developed a more ambitious and controversial project (*Nature* 396, 395; 1998), giving one company, deCODE, rights to its genetic data.

Latvia, with a population of 2.4 million, has a very small human-genetics research programme. The new initiative would allow this to be extended, as well as developing gene associations that might interest drug companies, says Elmars Grens, a molecular biologist at the University of Latvia in Riga, who helped to instigate it.

Grens is also co-founder of a start-up company called GENDP, which he hopes will win the contract for the high-throughput genotyping. The law has the support of the Latvian Academy of Sciences. Its vice-president, Juris Ekmanis, says it will mean a lot for Latvian science.

Grens says that a three-year pilot study, recruiting 40,000 patients, could start in 2002 if the law is passed, as many expect, by the end of this year.

Kári Stefánsson, director of deCODE Genetics, the company that controls much of Iceland's national genetic information programme, welcomed the Latvian initiative, but warned that "these initiatives are not easily turned into business". ■

Physicist claims gagging over missile defence system



Indefensible: Postol, seen here with a model of a mobile missile launcher, says the system is ineffective.

Jonathan Knight

A physicist and expert on national security says that the Pentagon is trying to silence his criticism of the US missile-defence programme.

Theodore Postol, Professor of Science, Technology and National Security Policy at Massachusetts Institute of Technology (MIT), has said for years that the technology required to create an effective missile shield over the United States does not exist.

The Defense Security Service, which handles military clearance for the Pentagon, has now asked MIT to recover the documents used by Postol to support his claims, and to persuade him to stop discussing them.

Postol has security clearance because he is a frequent adviser to the Pentagon on national-security issues.

Last year, Postol accused the Ballistic Missile Defense Organization (BMDO) of doctoring the results of a series of tests on the Pentagon's prototype missile-defence system to make it appear to be effective. In a letter to the White House in May 2000, he wrote that the BMDO's own data show that the system is less accurate than random guessing at telling real missiles from balloon decoys.

As evidence, he provided an unclassified report by the Department of Defense, detailing the BMDO test results. He also disseminated the report over the Internet.

But soon after, the Pentagon reclassified parts of the report as secret and sent three agents to discuss the issue with Postol at his

office at MIT. "I told them my policy is if anybody asks me for the report I give it to them," Postol says of the meeting. Earlier this year he provided it to the General Accounting Office, an investigative branch of the US Congress.

According to Postol, who now stands to lose his security clearance, the Defense Security Service's request to MIT is a clear attempt to keep him quiet. He points out that the documents are freely available on the Internet.

Many academics who specialize in government or military affairs have security clearance, and if the government can prevent public information from being discussed merely by declaring controversial material secret, says Postol, it could have a chilling effect on scholarship. "There are profoundly important issues of academic freedom at stake here."

The Pentagon has declined to discuss the case, but argues that holders of security clearance are required to keep secrets safe no matter how they were obtained, and even if they were obtained before they were made secret.

MIT President Charles Vest says that "MIT defends the right of our faculty to serve as responsible critics within the limits of the law". He did not say what action, if any, the institute will take. But he expressed concern in a statement that the government would try to classify widely available information as secret in order to limit the debate by academic experts. ■

♦ <http://www.armscontrol.ru/start/archive/news0005.htm#postol>

STEVE RAYNER/CORBIS



Capital gains? The Latvian capital Riga hopes to benefit from the information initiative.

Weizmann finance chief embezzled \$5 million

Haim Watzman, Jerusalem

An official at the Weizmann Institute of Science has confessed to pocketing US\$5 million from one of the institute's fund-raising organizations in Europe.

Israeli police say that some of the money, taken from the Zurich-based European Committee of the Weizmann Institute of Science, was channelled through the GSI, a national research institute in Darmstadt, Germany.

The Weizmann Institute in Rehovot is Israel's leading multidisciplinary research institute. It called in the police to investigate the activities of its finance division chief, Michael Netzer, in July. An employee at Bank Leumi in Zurich, where the European committee has an account, had notified the bank's branch in Rehovot of suspicious financial transactions.

According to a police spokesman, Netzer said that he invested \$1.5 million of the money in the stock market and used most of the rest to fund a gambling addiction. Netzer, whom the police have decided not to detain because of ill health, claims that he also used money to pay for medical procedures he needed.

The fund-raising European committee is one of several 'friends of' organizations that raise money for the Weizmann Institute around the world. Government funding covers only part of the institute's annual outlay,



Out of pocket: the Weizmann Institute called in the police to search for its missing millions.

and much of the remainder comes from funds raised overseas.

The European committee also holds an account in the Rehovot branch of Bank Leumi, where Netzer was one of four people with signing authority.

A subsequent audit of the Rehovot account revealed that Netzer deposited some of the money, which he claimed was 'private', into the personal account of the GSI's administrative director, who in turn made cash payments to Netzer. The administrative director was not available for comment, but Walter Henning, the scientific director of the GSI, says he sees no reason to conduct an

investigation at the institute.

Little information is emerging from the Weizmann Institute itself. Yivsam Azgad, an institute spokesman, has confirmed that the institute started the police investigation at the request of the European committee, which is legally an independent body.

Azgad states only that "a small part of the committee's funds were transferred to unidentified accounts" and that "a significant portion of these funds have already been returned". He added that the institute's investigation of the matter revealed "further irregularities", in the wake of which Netzer resigned from his post. ■

Public library set to turn publisher as boycott looms

Declan Butler

The Public Library of Science (PLS), a grass-roots initiative that has called for a boycott of scientific publishers, says that it is considering publishing journals itself.

Announced only three weeks before the boycott is due to begin, this is the next step in the library's campaign to promote free access to literature.

More than 25,000 scientists signed the PLS open letter, in which they pledged to stop buying, publishing in, or reviewing for any journal that refuses to place its research papers in free online archives six months after publication.

Michael Eisen, a geneticist at the University of California, Berkeley, and one of the leaders of the PLS initiative, says that one of the main goals of the campaign was to stimulate publishers into experimenting with business models other than the traditional 'reader pays' arrangement.

But most publishers seem uninterested, says Eisen, and the PLS has therefore "somewhat reluctantly" concluded that it will



Access all areas: Michael Eisen says he wants a 'positive option' for online science publishing.

probably need to create its own publishing system, to show the economic viability of alternatives and to provide the scientists who have supported its campaign with a place to publish that provides free access.

For the moment, the PLS has few detailed plans about how it would put such a scheme into practice, but says that it has discussed the idea with many of the signatories of the

open letter with a view to soliciting reviewers. Under its initial proposals for the scheme, all journals would be published online only, with page charges and institutional charges covering the estimated cost of US\$200–500 per manuscript.

With the September boycott deadline looming, the PLS is under pressure to maintain the movement's momentum. "A boycott is a negative thing. People have careers and need to publish; we want those who have supported the initiative to have a positive option supporting a new publishing venture," Eisen says.

The PLS does not expect that signatories will adhere to the boycott pledge to the letter. "People didn't sign to be dogmatic, but they share a goal and will back these up with reasonable action," says Eisen.

Although running close to its deadline, the PLS is open to discussing ways of moving forwards, Eisen adds. "A slight delay of a month or two in the boycott would be okay; any longer and we will lose the momentum." ■

Italian government promises to double researchers' money

Rome The Italian government will double its spending on research over the next five years, education and research minister Letizia Moratti said last week. She did not give details of how this would be achieved.

Moratti, who took up office in May, also announced her intention to "internationalize" Italian universities by attracting more foreign students and staff, and to modify the process for recruiting teaching staff. Under the current system, two or three successful candidates are identified by selection panels. Those not appointed are placed on a list from which other universities can appoint them without holding a competition. The list is to be abolished under the new proposal.

Government funding for Italian science currently stands at 1% of GDP, roughly half the European Union average. This lack of investment, together with the excessively bureaucratic recruitment process, has been blamed for Italy's comparatively poor standing in international science (see *Nature* **412**, 264–265; 2001).

Bioweapons labs 'could unleash forgotten diseases'

Washington Diseases such as smallpox could re-emerge as a result of research into defences against biological weapons, researchers cautioned last week (Srinivasan, A. *et al.* *N. Engl. J. Med.* **345**, 256–258; 2001).

Their warning comes after a US microbiologist was diagnosed with glanders, a contagious and potentially fatal disease that has not been reported in the United States for more than 50 years, though it is common among animals in developing countries. The researcher at the US Army Medical Research Institute of Infectious Diseases in Fort Detrick, Maryland, became ill after studying *Burkholderia mallei*, the bacterium that causes glanders. The institute works on defence strategies against biological weapons and naturally occurring diseases that require special containment.

The scientists warn that as research into bioweapons defence increases, so does the risk of laboratory exposure. "This case may serve as a harbinger of the resurgence of nearly forgotten diseases," they write.

Lucent losses arouse fears for Bell Labs

Washington The gloom surrounding Bell Laboratories deepened last week as Lucent Technologies, the telecommunications company that operates the labs, announced financial results that exceeded the worst



Ring of confidence? Lucent managers hope to avoid forced redundancies among researchers.

predictions of many industry analysts.

Bell Labs' headquarters in Murray Hill, New Jersey, is regarded by many as the world's premier centre for high-technology research, but recent financial difficulties at Lucent have raised fears that research there may be cut back.

Managers at Murray Hill are optimistic that there will be no forced redundancies among the 3,000 research staff, despite the net losses of \$3.2 billion that Lucent recorded for the second quarter of 2001.

The company is intending to shed 15,000–20,000 jobs from its workforce of 76,000; some 19,000 positions have already been lost since the start of the year.

Ötzi researchers uncover a stone-age murder mystery

Munich Sherlock Holmes might have solved the mystery a little quicker. Numerous scientists have pored over the body of 'Ötzi', a stone-age man who was discovered in the Alps in September 1991. But only now have researchers concluded that he was shot. Previous workers missed a crucial piece of evidence: the arrowhead is still in the body.

Conservators at Ötzi's final resting place, the South Tyrol Museum of Archaeology in Bolzano, Italy, announced that an X-ray analysis had shown the arrowhead — 21 mm long and 17 mm wide — to be lodged under his left shoulder. Computer tomography had also revealed a 2-cm hole in his shoulder



Iced: an arrowhead gives a clue to the reasons why Ötzi's body was found unburied.

blade, which could have been left by the arrow. Museum officials said that the possibility of a shooting had never occurred to other scientists who examined the body.

Lure of research rights entices workers back

Paris Staff at the Lure synchrotron at Orsay near Paris have ended a three-week strike after the French government agreed last week to grant them public-sector status when they move to a new facility in 2005.

The new synchrotron, known as Soleil, is due to open in 2005. The French government had asked CNRS, France's main agency for basic research and the majority shareholder in both Lure and Soleil, to set up a private company to run the new facility. This prompted Lure workers to down tools, claiming that commercial contracts would deny them the rights of researchers in the public sector, such as access to the research review procedures used to evaluate other CNRS research (see *Nature* **412**, 263; 2001).

Soleil will still be run by a private company, but the government has agreed to set up a special unit of the CNRS at the facility to employ the researchers.

Police foil two attempts to smuggle uranium-235

Paris Two apparently unrelated attempts to smuggle uranium-235 have been foiled by police in Europe.

The first, uncovered in Paris on 16 July, involved only 5 g of the isotope. But according to initial reports, the uranium was enriched at 80% — suitable quality for a nuclear weapon. Paperwork uncovered at the suspect's home indicates that the uranium came from an eastern European source. In a second incident, 1.7 kg of enriched uranium-235 was seized in a hotel room on 18 July in Batumi, Georgia.

Both uranium samples are now being analysed to determine whether or not they are highly enriched.

It is still rare for highly enriched uranium to find its way onto the black market. The Washington-based Nuclear Control Institute, a non-profit organization that monitors nuclear activities worldwide, says that 12 kg of uranium-235 is enough to make a dangerous nuclear device.

Correction

The photo legend in "Institutes prepare for pioneering bioinformatics work" (*Nature* **412**, 106; 2001) incorrectly stated that Nadia Rosenthal will lead the mouse biology programme at the European Mouse Mutant Archive (EMMA). The programme is in fact at the European Molecular Biology Laboratory (EMBL) in Monterotondo — EMMA is located on the same campus but is not part of the EMBL.

When the going gets tough ...

Britain's epidemic of foot-and-mouth disease gave the government's chief scientific adviser, David King, a baptism of fire. He has emerged with his political standing enhanced, says David Adam.

Starting any new job can be difficult, but spare a thought for David King, the British government's chief scientific adviser. Barely settled in behind his desk, King was thrown headlong in March into a crisis that threatened the future of British farming. Foot-and-mouth disease was spreading like wildfire, and epidemiologists were warning of a catastrophic "meltdown".

King seized the initiative. He surrounded himself with experts, including leading epidemiological modellers. This group's conclusions fed straight into government policy, through King's meetings with Prime Minister Tony Blair, held three or four times a week. In effect, King had wrested control from the Ministry of Agriculture, Fisheries and Food (MAFF). Almost daily, he faced the media to defend the government's handling of the crisis. "I was completely taken over by foot-and-mouth disease and my diary just had to be swept aside," he says.

For a British chief scientific adviser to have such a direct and obvious influence over government policy is unusual — King's predecessor, Robert May, never found himself on national television night after night. "Most politicians at least know who David King is now," says an official with one of the government-funded research councils. "Some didn't know who Bob May was even after he'd been there for five years."

For many, the pressure of becoming a pivotal figure in a national emergency would be hard to bear. But now the epidemic has tailed off sufficiently for King to grant an in-depth interview, it is clear that he thrived on the challenge. And he seems determined to use his raised political profile to maximum effect. King has already won a review of public spending on research, and now aims to improve the effectiveness of scientific advice across government.

There is "no question", King says, that the epidemic helped him to find his feet. More importantly, he thinks it showed politicians the value of scientific advice. "It has been enormously powerful in bringing to the attention of the prime minister how science can advise policy-making and be effective."

King came to his present role from Cambridge University, where he headed the chemistry department. He was born in South Africa in 1939 and graduated in chemistry from Witwatersrand University in



Power shift: David King (above) took the lead from the agriculture ministry in advising Tony Blair (inset) on foot-and-mouth.

Johannesburg, where he completed a PhD before moving to London's Imperial College in 1963. Spells at the universities of East Anglia and Liverpool followed, before King arrived in Cambridge in 1988. He has an international reputation for his work on surface chemistry and catalysis, and still carries out research in Cambridge for one day each week, supervising more than a dozen post-graduate students.

Key players

Officially appointed in October last year, King says he only completed the transfer from academic life at the end of December. In the weeks that followed, he concentrated on building his relationships with the key players on Britain's science policy scene. Then foot-and-mouth took over.

Reflecting on the handling of the epidemic, King says his only regret is his initial reluctance to get directly involved. "My first reaction was that this was something that wasn't going to go away quickly — foot-and-mouth never does," he says. "But I did feel that I would not interfere with MAFF and I

would leave them to run the campaign."

That changed on 21 March, when King attended a meeting with John Krebs, head of the Food Standards Agency, and four research groups that were using computer models to predict the spread of the disease (see *Nature* 411, 977; 2001). The epidemiologists all delivered the same, bleak message: foot-and-mouth was running out of control, and a more rigorous culling policy was needed to rescue the situation. "It became apparent that modelling was required," says King. "Then once you've got the models going you could set up different control scenarios and see what was needed."

King set the modellers to work. Within two days, he presented Blair with what became known as the '24-/48-hour' policy — to kill infected animals within 24 hours and those on surrounding farms within 48. King says that all four modelling groups predicted that such drastic measures were required. "So I felt fairly confident and stuck my neck out."

King makes no attempt to disguise his blurring of the line between advising on policy and making it. "When I presented my

JOHNNY GREEN/PA; INSET: CHRIS ISON/AP

policy decision to the prime minister there was a very short timescale," he explains. At that time the number of cases was doubling every nine days. "So the first thing I said was 'act fast,'" King continues. "We had calculated a whole range of scenarios but I simply said that this is the one that will work. So it wasn't a matter of giving what I thought would be a confusing set of options."

King is clearly comfortable with taking responsibility for the policy that was adopted. But as the cull passed 3 million, and images of mass graves and pyres dominated the media, criticism mounted, with farmers and some veterinary scientists questioning the need to kill so many animals (see Correspondence, page 477). Some critics in the media even suggested that the policy was designed to bring the epidemic under control before 7 June, the date of the general election, already delayed from 3 May by the epidemic.

King dismisses this idea, and plays down the significance of the disagreements between veterinary scientists and epidemiologists — both represented on his expert group. "It's certainly true that there has been a creative tension throughout, but I wouldn't say that there's been an argument between the vets on one hand and the epidemiologists on the other," he says.

King is adamant that the culling was justified. Some of the critics, he argues, do not understand the sophistication of the models that his group employed, and the fact that they were crunching 'real' data. "The outbreak had been going almost a month and by using the way it was developing you could tell what the incubation time and the infectious period was, and add that to the data to predict how it would go." King claims that had he not intervened, the eventual size of the cull needed to stamp out the disease could have been far worse. "We were predicting that we could lose half the farm animals in the UK."

Even with rigorous culling, the models predicted a long 'tail' of infections that is still occurring. King's main current worry is that complacency will set in, causing a relaxation in the 'biosecurity' measures introduced to prevent the disease's spread. If the outbreak



Killing fields: sheep waiting to be culled in Cumbria at the height of the foot-and-mouth outbreak.

survives the summer, he warns, it could flare up again, under the cooler conditions that favour the virus's transmission.

MAFF, meanwhile, did not survive the summer, being replaced after Blair's election victory with DEFRA, the Department for Environment, Food and Rural Affairs (see *Nature* 411, 727; 2001). Many observers feel that the defunct ministry failed to learn from previous mistakes. "The first lesson from the foot-and-mouth epidemic is the same lesson we learnt from BSE," says King. "That is: when there's a crisis you must bring on board as quickly as possible the widest range of scientific advice." Which MAFF failed to do? "Which I did," he says, with a relaxed smile.

What Tony wants

King, who claims to have received only a few hours of media training, certainly has the poise of his new political colleagues. At least at present, he also has regular access to the prime minister — which is of huge significance in a government said by insiders to be dominated by two words: "Tony wants ..."

How King uses this influence will be crucial. Already, he has ensured that science will be included in the next cross-departmental review of government spending, due to be completed by summer 2002. Two previous reviews, carried out since the Labour govern-

ment came to power in 1997, resulted in increased spending on research.

King's next priority is to improve the standard of scientific advice across the whole of government. On this point, he is not afraid to criticize his political paymasters, arguing that the decision to shed much of the government's in-house scientific expertise has caused problems. Although this policy was instituted by the previous Conservative government, Blair's administration has continued the trend, most recently spinning off three-quarters of the Defence Evaluation and Research Agency into a company called QinetiQ (see *Nature*, 412, 9; 2001).

"The net result is that we've lost a fair amount of the science base from within the civil service, so we don't have these people bubbling up into top positions," says King. "The problem with an organization that outsources all of its resources is that, if it does it too rigorously, it no longer knows even what questions to ask."

To plug these gaps, King wants to bring into government top scientists from universities and industry, starting at the successor to the ill-fated MAFF. "I expect that DEFRA will shortly be advertising for a chief scientific adviser along exactly these lines," he says. "They will parachute in an expert on environment and agriculture."

King admits that he is still getting used to his job, saying there is a "massive culture gap" between life as a university professor and as a civil servant. Maybe so, but King shows every sign of becoming an effective civil servant who is also a consummate politician. If that means that science assumes a central role in government thinking, he is likely to be a popular figure among British researchers.

"My belief, and the prime minister agrees with this, is that science today deals with exceptionally complex phenomena but can maintain a quantitative understanding," says King. "That was one of the lessons of this outbreak, that you can take a complex phenomenon and can really say something useful." ■

David Adam is a news and features writer for *Nature*.



Clear choice: with the weight of epidemiological data, King is adamant that the culling was justified.

Measuring the immeasurable

When chemists developed techniques to peer into the heart of chemical reactions, they opened up a new world for study. Yudhijit Bhattacharjee finds out how, a decade later, the early promise has been realized.

According to the textbooks, chemical reactions are a simple business. Take some reactants, mix, supply energy if required and, hey presto, the products of the reaction are created. But when it comes to details of the heart of the process — the fleeting moments in which individual reactant molecules turn into products — the textbook accounts were, until recently, a little vague. Everything happens so quickly, often in around 10^{-12} seconds, that chemists simply could not follow the action.

The development in the 1980s of lasers capable of delivering ultrafast pulses of light changed everything. Thanks in large part to the work of Ahmed Zewail at the California Institute of Technology in Pasadena — who used the new lasers as 'cameras' to study the intermediate stages of chemical reactions — the first snapshots of reactions with exposure time measured in femtoseconds (10^{-15} s) were taken soon afterwards. The field of femtochemistry was born.

As the technique becomes more widely used, a wealth of findings is emerging — from detailed analyses of photosynthesis to slow-motion replays of melting semiconductors. And chemists are not resting on their laurels. The next frontier — attosecond (10^{-18} s) measurements of the movements of electrons — is already in sight, offering chemists new insights into why some reactions occur readily, whereas others do not.

Zewail made femtochemistry possible by updating an existing chemistry technique, known as optical spectroscopy. First, the substances being studied are blasted by a short pulse of laser light that contains just

enough energy to kick-start the reaction. A second pulse is then used to probe the system while the reaction is taking place. This pulse passes through the reacting substances and is analysed as it leaves. Every atom or molecule has a unique set of frequencies of light that it can absorb, so by looking for frequencies that are missing from the pulse, chemists can determine what chemical species are present, and in what quantities. Atoms or molecules can also re-emit light that they have absorbed, again at specific frequencies.

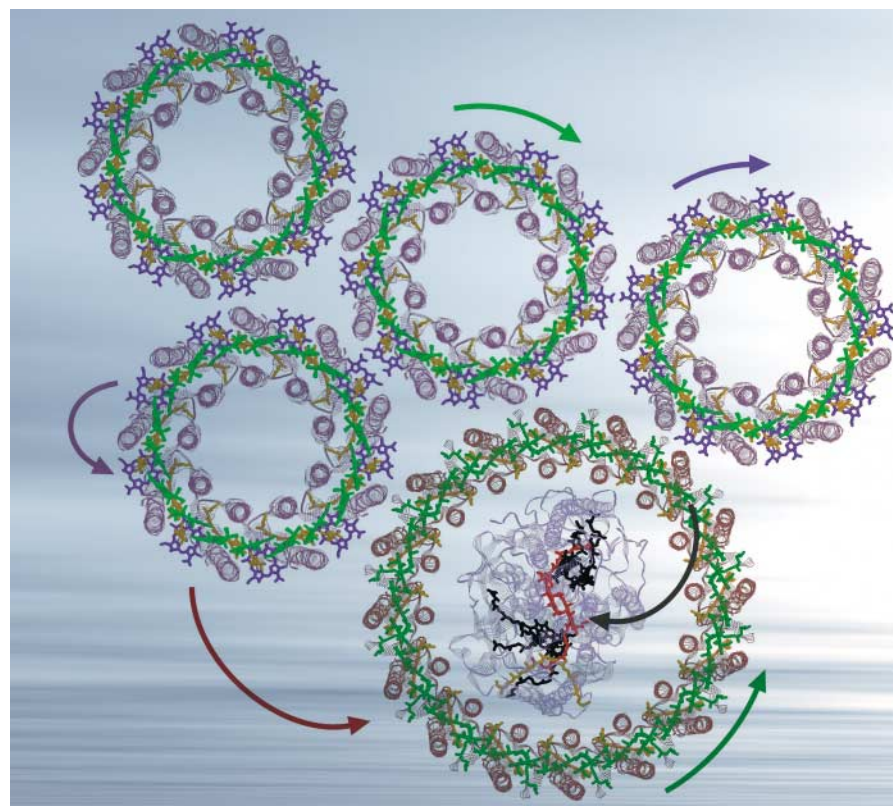
Pulse rate

By studying the absorbed or emitted light, chemists obtain a snapshot of a reaction taken as the probe pulse passes through the sample. Just as photographers use short exposures to capture fast-moving objects, so chemists need short laser pulses to study rapid reactions. A reaction can be over in 1,000 femtoseconds, so lasers capable of emitting pulses lasting a hundred femto-

seconds or less are needed to probe the details of the process.

In the 1970s, work by Charles Shank and colleagues at Bell Laboratories in Murray Hill, New Jersey, on liquid dye lasers, led to devices capable of producing pulses of just a few hundred femtoseconds. Researchers building lasers face a trade-off — the shorter the pulse, the larger the range of frequencies of light it must contain. Liquid dye lasers use fluorescent organic dyes to produce laser light. By using a range of dyes, each of which emits light of a different frequency, Shank and his colleagues were able to restrict the pulses to the very short times necessary. By the early 1980s, they had developed the technology so that pulses of tens of femtoseconds were possible¹.

In 1987, Zewail built a version of the Bell Labs laser that was capable of producing pulses of 60 femtoseconds and used it to study the dissociation of iodine cyanide (ICN) into iodine and cyanide². ICN molecules absorb different frequencies of light as the bond



Leading light: femtosecond spectroscopy was used by Graham Fleming (left) to follow the energy path of photosynthesis from the small light-harvesting complexes to the reaction centre.



TIFFANY DRESSEN

X. HU ET AL. PROC. NATL. ACAD. SCI. USA 95, 5935-5941 (1998)



Fast exposure: Ahmed Zewail used lasers as 'cameras' to study chemical reactions.

between the iodine atom and cyanide molecule stretches. Zewail took a series of snapshots of the process by sending in probe pulses at different times after the initial pulse and showed that the bond broke after around 200 femtoseconds, by which point the iodine atom and cyanide molecule had been stretched by about 5 angstroms. In itself, that finding was of limited interest, but the proof that femtosecond spectroscopy worked was a huge breakthrough. Twelve years later, Zewail was awarded the Nobel prize in chemistry.

More recently, femtosecond spectroscopy has been used to explain the remarkable efficiency of some photosynthetic reactions. The bacteria *Rhodospseudomonas viridis* and *Rhodobacter sphaeroides* interest chemists because they convert almost all of the light they absorb into chemical energy. Graham Fleming of the University of California, Berkeley, has helped piece together a detailed understanding of the chemistry that underlies this phenomenal efficiency.

The key unit in the collection of proteins that controls photosynthesis is a group of molecules, the reaction centre, that drives the reactions that convert energy from the Sun into the chemical energy needed to create organic matter. It is surrounded by a network of proteins, consisting mainly of chlorophyll. These proteins harvest sunlight, soaking up its photons and channelling their energy towards the reaction centre.

It is this channelling that fascinates Fleming and other chemists. The process starts with the absorption of a photon of light by a chlorophyll molecule. The chlorophyll gains energy from the photon, but this extra energy is rapidly transferred to a neighbouring chlorophyll molecule, which passes it on in turn. In this way, the photon's energy hops between chlorophyll molecules until it reaches the reaction centre.

Chemists suspected that the efficiency of the bacteria was due, in part, to the speed at which the energy is transferred between chlorophyll molecules. If energy is not passed on rapidly, it risks being lost as heat. Any pauses in the energy's journey to the

reaction centre would reduce the efficiency of photosynthesis.

The traditional model of how the energy is transferred, formed over half a century ago, predicted a much slower transfer than would be needed. Fleming developed the theory to produce an alternative version, complete with new predictions for the speed of the transfer. The predictions seemed plausible and, thanks to femtosecond techniques, Fleming was actually able to test them.

As the energy passes through the network of chlorophyll molecules, it changes the way they absorb and emit light. This allows the flow of energy to be detected by spectroscopy. Fleming followed the movement of the energy as it sped towards the reaction centre and showed that the jumps between molecules typically occurred in a few hundred femtoseconds³, just as his model had predicted. "Without femtosecond pulses," says Fleming, "we would not have been able to catch the excitation as it moves from molecule to molecule."

Energy drain

Fleming, together with plant biologist Krishna Niyogi, also at the University of California, Berkeley, is now tracking the flow of energy under different lighting conditions. In bright sunlight, plants reduce their light-harvesting efficiency to protect against over-excitation of the reaction centre, which can damage proteins vital to photosynthesis. Last year, Niyogi identified a protein, known as PsbS, involved in this regulation⁴. The Berkeley researchers are now using femtosecond techniques to test the idea that this protein performs its regulatory role by draining surplus energy from the chlorophylls and dissipating it as heat.

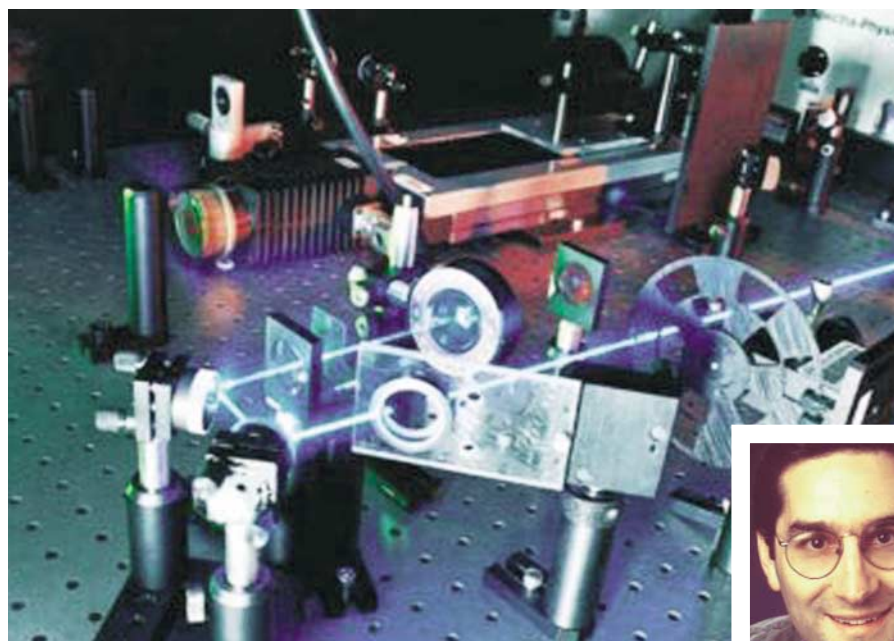
The ease and efficiency with which photo-

It's like following a hummingbird's wings — the movement is so rapid you need a fast camera.

synthetic systems channel energy towards the reaction centre has led some chemists to attempt to create artificial analogues. Researchers hope that the chemical energy generated by such a system could be harnessed to drive the synthesis of useful organic molecules. Just as naturally occurring photosynthesis was the first stage in the formation of oil and coal, artificial photosynthesis could help produce organic molecules for use as fuels.

By using long chains of polymer as light-harvesters, Benjamin Schwartz and Sarah Tolbert of the University of California, Los Angeles, have taken the first steps to developing such an artificial system⁵. The polymers were aligned in fine silica channels with their ends, which stuck out of the channels, grouped together into a bundle. Schwartz and Tolbert illuminated the bundle with visible light and used femtosecond probe pulses to follow the energy absorbed by the ends of the chains as it was channelled towards the section of the chains in the silica.

"It will be many years before we've successfully reproduced what has taken nature a billion years to design, but I do believe we will eventually be able to achieve systems that are as efficient," says Schwartz. "The time-resolved experiments are the key to determining exactly where the energy is



First light: Benjamin Schwartz (inset) and colleagues have used femtosecond lasers to study artificial analogues of photosynthetic systems.

UCLA

going and how long it takes to get there."

Other researchers are using femtosecond techniques to reveal how genetic material protects itself against damage by ultraviolet light. Last year, Bern Kohler of Ohio State University in Columbus described how he dissolved DNA and RNA bases in water and used pulses of ultraviolet light to increase their energy⁶. He then used probe pulses of visible light to measure how long they took to fall back to their ground states.

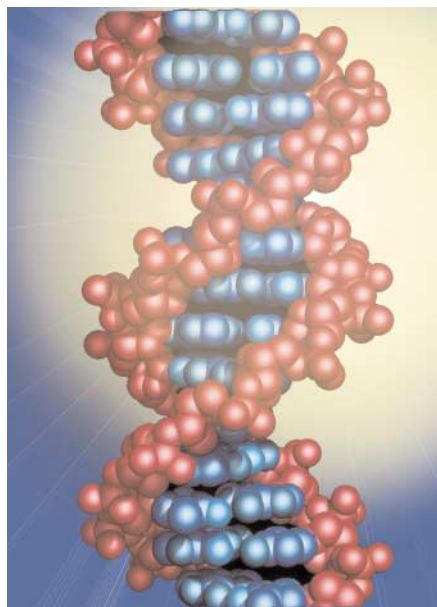
Chemists knew that the bases were able to return to their ground states very rapidly but had never studied the precise timing of the process. They suspected that the speed of the jump was one reason why nature had selected these molecules over others. Bases are more likely to react with other substances when in an excited state, so the quicker they return to their ground state, the smaller the chance of them reacting and damaging genetic material.

Kohler found that the bases do indeed lose their energy very rapidly — taking between 290 and 720 femtoseconds to do so. He also revealed that this speedy jump was made possible by the bases using the extra energy to heat the surrounding water, rather than emitting electromagnetic radiation, which is orders of magnitude slower than heat transfer.

"In a way, DNA bases serve as their own sunscreen," Kohler says. "Photostability would have been all the more critical when life first appeared on Earth, because there was no significant ozone layer in the atmosphere to shield against ultraviolet radiation."

Visible light is not the only kind of probe pulse that can be used. Light provides only limited information on the structure of a substance, so some researchers use femtosecond X-ray bursts as the second pulse. X-rays usually carry too much energy to excite the substance being studied. Instead they pass through the gaps between atomic nuclei. Because the wavelength of X-rays is similar to the size of these gaps, the waves are diffracted. In crystals, the regular spacing of the atomic nuclei causes interference patterns in

Femtochemistry techniques have helped to reveal new classes of reaction.



Sunblock: DNA bases are their own sunscreen.

the diffracted rays. Researchers can work backwards from these patterns to determine the structure of the molecules involved.

Such techniques have helped reveal classes of reaction that researchers had previously been unaware of. Ultrafast melting is one example. When a laser pulse strikes the surface of a semiconductor, electrons in atoms at the surface absorb energy from the pulse. This energy is transferred to the thermal motion of the atoms over a period of a few thousand femtoseconds. Melting was thought to begin when the thermal energy was such that the atoms' vibrations broke the bonds between them.

But in 1999, Arlin Chin of the University of California, Berkeley, melted a semiconductor with a pulse of laser light and showed that melting began before the thermal energy had transferred to the atoms⁷. Chemists were at a loss to explain how the atoms had broken the bonds between them without gaining the thermal energy that is usually needed to do so.

Earlier this year, researchers led by Antoine Rousse of the Laboratory of Applied Optics in Palaiseau, part of the CNRS, France's national research agency, extended Chin's work. They showed that the atoms began to lose their regular structure in as little as 350 femtoseconds after the initial pulse was delivered⁸ — well before the atoms had gained the thermal energy needed to melt. The process seemed to be occurring because the electrons that absorb the initial laser

pulse can move away from the atom they were originally associated with. This reduces the forces that keep each atom fixed in position. As the atoms start to move more freely, the semiconductor enters the liquid state.

"It's like following the wings of a hummingbird," says Rousse of the femtosecond techniques he employed. "Your eyes can't follow a movement that rapid, so you need a fast camera. But ultrafast melting is 10 billion times faster yet."

Speed demons

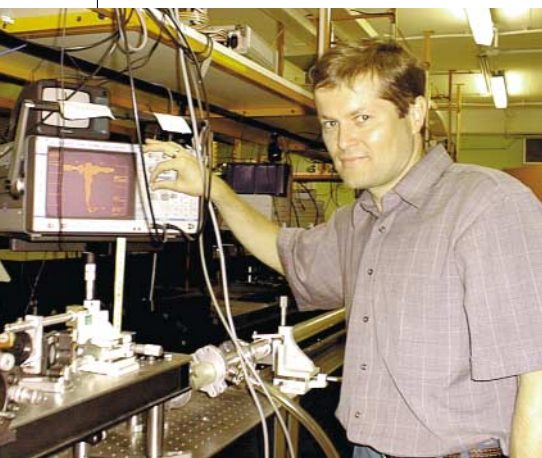
Despite the wealth of discoveries made possible by femtosecond techniques, some aspects of chemistry remain inaccessible. Whatever the probe pulse used, femtosecond techniques provide little information about the position of electrons within their orbits. According to classical theories, an electron can orbit a hydrogen atom in a fraction of a femtosecond, so pulses of just a few hundred attoseconds (10^{-18} s) would be needed to track each electron.

Some researchers believe attochemistry may soon be possible. This March, Markus Drescher and Ferenc Krausz of the Photonics Institute of the Vienna University of Technology described how they generated X-ray pulses of just 1.8 femtoseconds in length⁹. The lower limit for a pulse of visible light, defined by the duration of a single cycle of the wave, is 3 femtoseconds. But X-rays have shorter wavelengths, and pulses down to as little as 50 attoseconds are possible in principle.

If attosecond pulses can be generated, a slew of new phenomena will be open to investigation. The movement of electrons in excited molecules is one example. A molecule is more likely to react when one of its electrons is in an excited state. But the electron may fall back to its ground state before the reaction occurs. Studying the movement of excited electrons may help explain why certain reactions occur, whereas others fail to get going.

Attochemistry would be the latest in a line of advances that have redrawn the limits of the chemical sciences. Manfred Eigen, then director of the Max Planck Institute for Biophysical Chemistry in Göttingen, Germany, accepted the 1967 Nobel prize in chemistry with a lecture about his work on "immeasurably fast reactions". Zewail, who analysed reactions over timescales a billion times shorter than Eigen, later joked about this title. If attosecond techniques do become reality, the immeasurable will have to be redefined once again. ■

Yudhijit Bhattacharjee is a writer in Columbus, Ohio.



Faster still: Ferenc Krausz believes that attochemistry could soon be possible.

1. Fork, R. L. *et al. Appl. Phys. Lett.* **38**, 671–672 (1981).
2. Dantus, M. *et al. J. Chem. Phys.* **87**, 2395–2397 (1987).
3. Jordanides, X. J. *et al. J. Phys. Chem. B* **105**, 1652–1669 (2001).
4. Li, X. *et al. Nature* **403**, 391–395 (2000).
5. Nguyen, T. Q. *et al. Science* **288**, 652–656 (2000).
6. Pecourt, J.-M. L. *et al. J. Am. Chem. Soc.* **122**, 9348–9349 (2000).
7. Chin, A. H. *et al. Phys. Rev. Lett.* **83**, 336–339 (1999).
8. Rousse, A. *et al. Nature* **410**, 65–68 (2001).
9. Drescher, M. *et al. Science* **291**, 1923–1927 (2001).

Vets asked valuable questions about foot-and-mouth measures

Opposition was to a rigid policy, not to epidemiologists.

Sir — Your Opinion article “Lessons from an epidemic” (*Nature* **411**, 977; 2001), which identified an apparent rift between epidemiologists and some veterinary scientists, does nothing to analyse that rift. This is unfortunate: many veterinarians have been highly praised by the UK farming community. Surely you cannot criticize veterinarians for asking whether so many animals needed to be killed, or whether vaccination should have been used, or about the sensitivity of the epidemiologists’ models? These are complex questions directly involving the lives of our patients and the livelihood of our clients. It is surely our duty to question the veracity of policies that have such swingeing effects.

It is true that all disciplines must guard against myopia and welcome excellence from outside to advance their science. We have learnt the advantage of this in recent years: for example, following the United

Kingdom’s BSE epidemic, chairs in veterinary epidemiology have been created in the country’s veterinary schools.

The resentment to which you refer in your editorial, and which existed among some veterinary scientists, was not against the epidemiologists’ models but against the rigid policy being adopted by the government following advice from these scientists, which did not allow any flexibility for local veterinary assessment of risk. To slaughter animals when there is no risk is not only immoral and unscientific, it is bad veterinary medicine.

In some hotspot areas, the UK Ministry of Agriculture, Fisheries and Food (MAFF) — now the Department of Environment, Food and Rural Affairs (DEFRA) — has lost the confidence and support of local farmers, largely because so many animals have been slaughtered unnecessarily. No national animal disease control programme will succeed without the full

cooperation of the farming community.

There is plenty of evidence that veterinary scientists are and have always been willing to work with the wider scientific community, but, for successful communication, the dialogue must be in both directions. An example of where this did not occur may have been the basis for your editorial. On 23 March, a group of epidemiologists from Imperial College London did not endear themselves to any of the stakeholders in the foot-and-mouth crisis when they refused to release the details of their modelling work until it had been published on 12 April (*Science* **292**, 1155–1160; 2001). Fortunately, other epidemiology groups from Edinburgh, Cambridge and the Veterinary Laboratories Agency did make full details of their work and their opinions immediately available.

We accept that there are lessons to be learnt from this outbreak and we are keen to learn them. We are fully paid-up members of the wider scientific community and will continue to cooperate outside our profession, particularly in the control and prevention of animal disease.

R. G. Eddy

Royal College of Veterinary Surgeons, Horseferry Road, London SW1P 2AF, UK

Always a role for debate between disciplines

Sir — The Opinion article “Lessons from an epidemic” (*Nature* **411**, 977; 2001) fills me with concern for the future of science and public scientific debate. It would appear that you believe veterinary scientists should not be allowed to question the scientific conclusions reached by another discipline. The implication that these scientists are resentful at the leading role played by epidemiologists in the UK foot-and-mouth-disease crisis is misleading and untrue.

Surely there is always justification for scientific debate, questioning and investigation as to whether the output of the science is valid, sensitive and appropriate for the situation in which it is being used? In the case of foot and mouth it seems very appropriate for models based on mathematical formulae to be commented on by those involved in the control of disease. The questions posed — about the use of vaccination, the sensitivity of the epidemiological models and whether so many animals need to be killed — are of interest to the community at large.

Your Opinion article suggests that scientific questioning of those in one specialist area by those in another is not permissible, whereas in my opinion this is

quite justified and indeed follows the concepts laid down by Lord Phillips in the UK BSE inquiry. Equally, I have no doubt that models must be used even more in future, although again it must be recognized that there are various types of models producing different results for use in different circumstances. This requires that mathematical modellers, the veterinary profession and other advisers work together as a team to produce the best advice and a range of options. Economists also need to be included, as in many cases their role is equally important in determining the most cost-effective method for control of outbreaks such as the foot-and-mouth epidemic.

I echo your conclusion that various scientific disciplines must work closely together in future. Each will bring its own expertise, the sum of which will produce far better results than each alone, and enable those determining policy to receive the best advice. In the case of foot and mouth, I facilitated the rapid provision of the necessary data and I would hope that we can build on this in the future.

Nevertheless, all scientists, whatever their discipline, have the right to question the conclusions of others.

J. M. Scudamore

State Veterinary Service, Department for Environment, Food and Rural Affairs, IA Page Street, London SW1P 4PQ, UK

e-access

The debate continues at <http://www.nature.com/nature/debates/e-access/index.html>.

Beneficiaries should pay

Sir — The debate about free access to publications in Commentary, Correspondence and your website (see www.nature.com/nature/debates/e-access/ and, for example, *Nature* **411**, 521–522; 2001) is valuable. I have not, however, seen taken into account the point that — despite the bias introduced by the publisher-perish ethos — research publication contributes an ‘external benefit’ to appointments committees, grant-awarders and those evaluating research careers, which could not be obtained in any other way, or, probably, at any price. This means that the activities of these organizations are indirectly subsidized by subscribers to science journals.

External benefits can and should be paid for. It would be a reasonable act of good faith if these committees, grant-awarders and other organizations were to make a contribution to an academic library every time they took a decision that depended necessarily on the journal system these libraries finance. Can a mechanism be found to implement this suggestion?

Henry Nathan

Scientia, 12 rue Jacob, 75006 Paris, France

Research doesn't denigrate humanity

Being human is more than simply having the right molecular composition.

Hubert Markl

Probably nowhere in the world is the debate about the nature of humanity, its rights and obligations, more heated than in Germany. The debate extends far beyond current concerns about stem-cell research and pre-implantation genetic diagnosis, to fundamental questions about the nature of humanity, the freedom of science and its mission.

Who could fail to recognize that we are living in a year of the life sciences? The BSE crisis; devastation to sheep and cattle by foot-and-mouth disease and its control; panic over genetically modified crops; threats to global biodiversity; HIV; the whole area of stem-cell research, therapeutic cloning and pre-implantation genetic diagnosis of hereditary diseases — these are enough to make some people long for the good old days of the century of physics!

We have witnessed the unparalleled spectacle of Francis Collins and Craig Venter climbing down from a mountain of data like Moses from Mount Sinai, holding in their hands stone tablets engraved with 3.2 billion nucleotides. Lo and behold, the people were celebrating the sequenced genome, with songs of rejoicing in the feature and opinion sections of every newspaper. Who cares if it had only been partially sequenced and — as it now turns out — is not even quite correct? The extent of biotheological exaggeration was limitless. Some said the handwriting of God had been deciphered. Well, if the Bible is the book in which man described God, then surely it is a short distance from the reverse, though erroneous, conclusion, that genes are the writing in which God (or nature) has formulated the essence of mankind.

Ethical education

Despite the exaggeration, biologists and doctors have seldom learnt so much about the basics of life in such a short time, or faced so many possibilities for using this knowledge. And never before has such a wide public noted that much of what is new and not yet understood affects our most personal decisions: from what we eat to having children, from life insurance to jobs. This fills us with insecurity as well as hope. Because even we scientists can hardly keep up with the pace of research progress, we have to take this worry seriously and show that we understand it via public debate on these biological innovations. The media cacophony is an unavoidable, even indispensable, part of this debate. By this process, many people will see that their anxieties are exaggerated, as may be their hopes.



The Rubicon does not have evil looming on the other side, for if anywhere, evil is right here with us.

We shouldn't be surprised that so many people are calling loudly for the authorities to give them clear direction and visible boundaries, anchored in eternal values and truths. Yet no ethics council, however careful, can spare us from educating ourselves on vital issues so that we can make our own judgments, for every human being is ultimately responsible for the decisions he or she makes. Those who wish to end the confused debates with well-intentioned words of moral or legal authority will discover that the debates will go on for that very reason.

This all boils down to the eternal question: "What is a human being?" Heated debate has already progressed to the extent that some biologists will soon not dare enter the laboratory or appear in public unless accompanied by a constitutional lawyer and a moral theologian, or both. Every human being is new, unique and developed from a fertilized egg cell. However, the fertilized egg is far from being a human being in the full sense of that word: it can be called a human being only if the word is given a meaning

totally different from its usual definition. When we refer to an organism as 'human', this is an expression of self-reference, the meaning of which is stipulated not by nature but by humans themselves. 'Human' is a culturally defined attribute, not a purely biological fact — we consider a person who has died still to be a human being.

What is a human?

Every living human today belongs to the species *Homo sapiens*. Yet we call our biological ancestors 'human' more or less arbitrarily, as we cannot test ancestors to see if they can reproduce, or tell by looking at a fossil to which genus or species it should be assigned. As Jared Diamond has so clearly pointed out, from a purely genetic point of view, *Homo sapiens* could be designated as the third species of a genus of chimpanzee. But another insight, which we owe to Ernst Mayr, is that *Homo sapiens* should be classified not merely biogenetically, but according to its emergent, cultural level of achievement. Though humans are rooted in their biology, they simultaneously tower above it. That is why humanness cannot be defined exclusively in molecular-genetic facts such as the 3.2 billion nucleotides arranged in a specific order inside a zygote.

Today, as in the past, different cultures take different approaches to deciding the point after fertilization when a developing embryo becomes a human being. Judaism, to take a convincing example, assigns special significance to the fact that a human can only

develop when in intimate contact with the mother's body. Assigning to a zygote before implantation in the womb, or an embryo not physically and psychologically accepted by its mother, a quality such as 'human dignity', means subjecting human embryos to different norms from those used for other species — mouse embryos for example — in terms of authorization for use in research. The quality of humanity has to be attached to a stage of embryonic development. Otherwise, the largely respected legal regulations regarding abortion and the generally accepted use of implantation-blocking drugs or devices as a means of birth control would not be possible.

At the end of last year, the British parliament voted by a large majority to allow research on human embryos and cell cultures made from such embryos, even including therapeutic cloning, in the first two weeks of life under strictly controlled, justified conditions. Similar regulations are imminent in France, and will probably soon follow in other European countries. Of course, there is nothing forcing us in Germany to follow suit. But perhaps it is not a bad idea to consider the careful arguments of other countries in our community before we and the Vatican alone take the high ground of final moral justification, as is supported by our president, churches, some political parties, bioethics councils or many sections of the mass media.

In bioethical decisions, a free society must value highly the conscience and action of every individual, whatever their religion. Parents, especially mothers, must be allowed to decide whether to carry a baby to full term when there is possibility of severe developmental disorders in the embryo. When others take it upon themselves to dictate such decisions, it is as though a woman and her reproductive capacity (and even disabled people themselves) belong first to society or to the state, which either grants or denies her freedom about the most private of human rights, that of deciding when and how to use her own reproductive capacity.

The older I get, the more I consider it wrong when old men such as myself try to force young men to go to war or young women to reproduce against their will. It may be true that parents have no legal entitlement to a healthy child, but they certainly have the human right to strive for one! Nor does a society have the legal right to force people to give birth to disabled babies as proof of its moral principles. I call for freedom: not necessarily for freedom of research — that is a minor, though valid, concern — but for the freedom of every citizen in a democratic society, which is under threat from the coercion of a zealous 'moral majority'.

When does an individual human life begin? Of course, the egg and sperm are both alive: they have genomes, each genetically is an individual. But the development, or

A human being is made not at conception but when the zygote becomes implanted.

epigenesis, of a new person is made possible by the formation of a zygote, not determined by it. The zygote has the potential to become a human being, but only under certain conditions in which, for mammals, being linked to the mother is not like the relationship between a landlord and tenant renting a biological apartment, but constitutive for normal development.

What is life?

Biologically, this fact has far-reaching significance. To begin a symbiotic relationship with an embryo is an individual decision full of tremendous consequences for success in life, because a mammal cannot simply abandon her embryo as a bird can leave her eggs. For long-lived species with only a few chances of producing offspring, such as ours, this decision is especially crucial to life. The 'biological' decision to create a human being is made not at conception but when the zygote becomes implanted in the uterus, which happens in only relatively few cases.

Embryos that spontaneously abort often have genetic anomalies. The most common anomaly due to triplication of a single chromosome occurring in humans is trisomy 21. Anyone tempted to suggest that the resultant developmental defect (Down's syndrome) is ordained by God has to accept that most other trisomies of single chromosomes also occur in humans, and must be just as ordained by God, but are spontaneously aborted.

It would, of course, be a naturalistic fallacy to take the natural selection in the womb of genetically defective embryos as justification for attempts to achieve similar effects by artificial selection. Nevertheless, humans, who have during our evolution towards reason been largely set free from genetic determinism, are entitled to do what nature would do anyway through intentional intervention if an individual considers it right. It seems to me that humanity's concentration on the possession of a set of human genes, and the fatalistic acceptance of every negative twist of fate as being in the make-up of these genes, is the ultimate in 'biologism', degrading us to simply a product of biology and depriving us of precisely that which makes us human: our culturally based freedom of choice.

Facts such as these should not be disregarded in the evaluation of pre-implantation genetic diagnosis of severe genetic disorders, which is sometimes misleadingly or mali-

ciously equated with the Nazi methods of selecting victims for the gas chamber. Those who do not recognize the difference between a parent's decision about an embryo or fetus, and the murder of children and adults to serve the supposed public interest, have the least chance of doing justice to the memory of the victims of Nazi state-sponsored terrorism.

The German president, Johannes Rau, was right to warn us scientists to uphold ethical values. Only last month, I apologized to the victims of biomedical and racist research madness during the Nazi era for the moral failures of the guilty members of the Kaiser Wilhelm and Max Planck societies, and I asked the victims for forgiveness. But we must categorically distinguish between the atrocities of scientists in a regime of terror, and the procedures used in research and medicine for pre-implantation genetic diagnosis, therapeutic cloning and development of treatments for serious diseases using cell cultures from embryos. Equating the one with the other is totally wrong and belittles the suffering of the Nazis' victims. Everyone agrees that these victims were misused and humiliated human beings, whereas there is no biological reason to attribute complete personhood to a few-celled embryo simply because, in interaction with a mother organism, it has the ability to become one — although its dignity should certainly be respected from its very beginning.

This is why, as a member of the senate of the DFG (Germany's main funding agency for university research), I voted in favour of its position on stem-cell research. The DFG's suggested course of action is moderate and the planned further research appropriate. I thank the president of the DFG, Ernst-Ludwig Winnacker, for standing firm despite personal insults and widespread denigration from the public.

Basic researchers seeking knowledge for its own sake have no need to experiment on human embryos, as they can use mice or other lab animals. It is the therapeutic goal of healing severely ill patients that must be considered by legislators when deciding whether to allow human embryo research. The vast majority of researchers desire to benefit individuals, society and the environment, and do not have delusions of grandeur or visions of creating a Frankenstein's monster. The Rubicon does not have evil looming on the other side, for if evil is anywhere, then it is right here with us and has been for some time. Instead, we must find new ways to navigate this river, and cross it only in the knowledge of our full responsibility for our actions. ■

Hubert Markl is president of the Max Planck Society, Hofgartenstrasse 8, 80539 Munich, Germany. This article is an edited version of a speech (full text on www.mpg.de) given by the author at the plenary assembly of the Max Planck Society on 22 June 2001.

Ecology goes macro

Taking a bird's-eye view reveals the hidden order of ecosystems.

Pattern and Process in Macroecology

by Kevin J. Gaston & Tim M. Blackburn
Blackwell Science: 2000. 377 pp. £39.99,
\$79.95 (pbk)

Pablo A. Marquet

Ecological systems are paradigmatically complex. They are composed of spatially and temporally intricate interacting networks of different entities, such as species, whose functioning generates the conditions for their own existence.

Ecologists have traditionally tried to understand ecological systems by working with a limited subset of species in a well-defined area, usually a small one, and characterizing their interactions through observation or experiments, usually of short duration. This strategy has revealed myriad mechanisms of direct and indirect interactions between species, but fails to provide an understanding of the dynamics of the whole system.

Unsurprisingly, under this view, ecological systems are pictured as highly idiosyncratic, and explanations of their functioning become contingent on the organisms present, and the particularities of the environment, space and time. Such contingency makes any generalization of experimental results venturesome, at best. However, such a reductionistic or 'microscopic' view can be fruitfully complemented with 'macroscopic' approaches, which sacrifice the detail and contingency of the local scale in the hope of identifying general principles or statistical patterns in ecological systems. In a paper published in 1989, James Brown and Brian Maurer dubbed this approach 'macroecology', which is the subject of Kevin Gaston and Tim Blackburn's book.

The authors aim to show that the macroecological approach is interesting and informative and to "promulgate an understanding of why it is such an important part of the broader programme of research into ecology". They start with a description of what the macroscopic perspective entails and end with a chapter attempting a synthesis of patterns and processes.

The book is organized along four



Creatures great and small: the jay (above), willow warbler (left) and green woodpecker occupy different woodland niches.



main axes of macroecological enquiry: species richness, range size, abundance and body size. The analysis of patterns, processes and explanation along these four axes is, for the most part, exhaustive and provocative.

Although I do not agree with several of their explanations and find their synthesis still rudimentary, their analysis will surely spur much research in macroecology.

To make their analysis accessible to a broad spectrum of ecologists, the authors chose the birds of Britain (an area of 230,000 km²) as an exemplary large-scale assemblage with which to explore macroecological patterns and process. They aim to illustrate how processes and patterns at this scale provide insights that can be essential for understanding the structure and dynamics of small-scale local assemblages, as exemplified by the birds found at Eastern Wood (0.16 km²) in the county of Surrey. This is probably the first com-

prehensive macroecological analysis of a local assemblage.

The choice of birds is not surprising, given the British enthusiasm for them and the enormous amount of data collected about them (the first complete census of Eastern Wood birds was made in 1949 and records run almost uninterrupted until 1979). What is so special about Eastern Wood, besides having well-watched birds? The answer is simple—nothing. That is precisely the point the authors want to make: The macroecological approach is applicable to any ecological system.

Gaston and Blackburn place much emphasis on showing that ecological systems are continuous across scales—that patterns and dynamics at local scales are not independent of processes operating at larger scales.

However, in making apparent the benefits of this approach, they also highlight its limitations. The first of these is that most of the macroecological patterns discussed in the book are of species richness, abundance, range size or body-size distributions at different spatial scales. This emphasis on spatial embedding relegates



to a secondary position the equally important role of processes occurring across time. These are manifest in historical contingencies that are known to affect macroecological patterns such as body-size distributions (the 'temporal embedding problem'). Second, because macroecology is mostly concerned with, although not limited to, the search for statistical regularities, the analysis of patterns is usually restricted to particular taxa for which large amounts of data are available. To what extent does knowledge generated for a particular taxon extrapolate to communities or assemblages of diverse organisms living in the same habitat?

Although it is wise to start the study of complex ecological systems in a simple way, I think macroecology should go beyond this taxon-based approach to the study of the ecological systems in which particular taxa are embedded. Resolution of this 'biotic embedding problem' requires more emphasis on the empirical analysis of macroecological patterns across taxa within communities and less on compilation studies for particular taxa.

Throughout the book, the authors correctly point out that macroecology rests heavily on the use of the comparative methodology — the description of patterns and the development and testing of hypotheses using information on the distribution and covariation of traits (such as abundance and body size) across species. In this kind of analysis, species are not independent realiza-

tions but are linked to each other by shared ancestry, which raises the statistical problem of non-independence of data. The authors acknowledge this, cautioning on the need to remove the potentially confounding effect of shared ancestry by using phylogenetically independent comparisons, which supposedly remove the phylogenetic component from the pattern.

This is a contentious issue, however, as the phylogenetic history of taxa is not independent of their ecology. To some extent, what we call history is the history of the ecological interactions between species and their environments, and that is reflected in extinction and diversification patterns through time. When phylogenetic signals are removed from the data, their ecological correlates are also removed. Further, even if we could assume that ecology and phylogeny are independent, too much emphasis is placed on removing potential phylogenetic effects on patterns and too little on actually quantifying how much variance in the pattern is explained by phylogeny and why it is stronger in some groups or for some traits than others.

Macroecology is in essence a discipline of synthesis, whose main aim is the search for general principles or natural laws underlying the seemingly endless variability of life in its many forms of organization. After reading Gaston and Blackburn's book, however, it is clear there is still a long way to go. Macroecology is becoming stuck in

contingent explanations for many of the large-scale patterns it addresses. However manageable this contingency could be, more effort should be directed to get beyond it.

The structure and function of present-day ecological systems can be regarded as the results of the unfolding process that started with the biotic Big Bang that was the emergence of life on Earth. The task of the ecologist is to disclose the hidden order and the rules that govern this process of unfolding. Gaston and Blackburn's book is an important step along this path — an authoritative characterization of the status of this research programme in ecology, a thorough pattern-by-pattern analysis and a provocative statement on what macroecology is and what we should expect from it. ■

Pablo A. Marquet is in the Department of Ecology, Universidad Católica de Chile, Casilla 114-D, Santiago, Chile.

What you see ...

Visual Disturbances following Gunshot Wounds of the Cortical Visual Area

by Tatsuji Inouye (translated by M. Glickstein & M. Fahle)

Oxford University Press: 2000 (German original published by Wilhelm Engelmann: 1909). 101 pp. \$25

Daniel L. Adams and Jonathan C. Horton

The cerebral cortex is divided into dozens of areas, each devoted to processing some element in the human repertoire, such as vision, hearing, touch or movement. A basic principle of neuroscience is that many areas contain an orderly, topographical representation of the function that they serve.

The first such map was made for the visual cortex by a brilliant young Japanese ophthalmologist, Tatsuji Inouye. On 8 February 1904, shortly after his graduation from Tokyo University, the Russo-Japanese war erupted. Inouye's duty was to assess visual loss in Japanese soldiers following brain injury so that their pensions could be adjusted suitably. Dissatisfied with this mundane task, Inouye set out to discover exactly how the visual world is represented in the brain. The resulting monograph was published in German in 1909. Unfortunately, only a handful of copies were printed, and Inouye left science soon afterwards to pursue medicine. His seminal contribution was lost, until Glickstein and Fahle provided this translation of a photocopy of the original, which they had discovered in the

The art of botany

A page from the sixth-century Byzantine *Codex Aniciae Julianae*, the oldest illuminated copy of the writings of Dioscorides, ancient botanist and pharmacognosist.

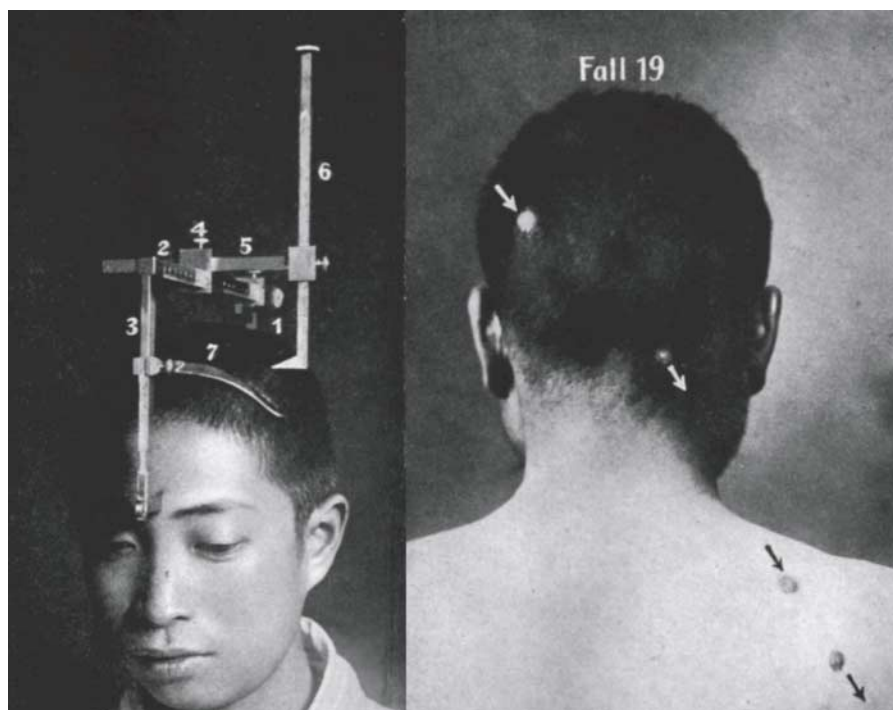
This is one of around 500 botanical illustrations, covering 15 centuries, in the trilingual book *Ein Garten Eden*, which accompanies an exhibition of some of the Austrian National Library's extensive collection.

The exhibition runs until 31 October 2001 at the Austrian National Library in Vienna.

Garden Eden: Masterpieces of Botanical Book Illustration by

H. Walter Lack (available in English, German and French; Austrian National Library, US\$39.99, £19.99, DM49.95, FF262.50).





The cranio-coordinometer (left), used to plot the trajectory of bullets through the head (right).

library of the Institute of Neurology in London.

To map the primary visual cortex, Inouye had to overcome the problem posed by individual variation in head size and brain anatomy. This variability still dogs researchers today, especially those who use functional imaging techniques. Inouye's solution was to compile an average head model by measuring cranial landmarks using a stereotactic instrument he invented called a cranio-coordinometer. Cadaver brains, cut sagittally, were photographed and projected onto the head model to establish the coordinates of the calcarine fissure, which contains the primary visual cortex.

The map was constructed by examining 28 soldiers (selected from 80,000 wounded) who had suffered bullet wounds to the occipital lobe. The site of brain damage was correlated with the defect in the visual field to compile a retinotopic map — the 'projection' of the retina on the visual cortex. The Russians used a high-velocity rifle, the Mosin-Nagant Model 91, that fired a 7.62-mm hard-jacketed bullet which pierced the skull without shattering it. It left tidy entrance and exit wounds, along a straight trajectory, that could be measured using a tape measure and translated onto the head model to plot the swath of destruction through the visual cortex.

Inouye showed an innovative flair and analytical rigour that were absent in his contemporaries. Among many insightful observations, his study contained the first accurate description of the arrangement of the visual field map in the primary visual

cortex. The fovea was placed correctly at the occipital pole, and the peripheral field at the anterior end of the calcarine sulcus. The upper and lower visual quadrants were mapped onto the lower and upper calcarine banks, respectively. Inouye illustrated the retinotopic map as a quarter-sphere, distorted into the shape of a trapezoid and projected onto the surface of the visual cortex. He believed that the distortion followed a mathematical function defined as the 'area-true representation'.

His map magnified the most central, foveal portion of the visual field. Selective magnification of maps is a common theme in the cerebral cortex. For example, in the somatosensory cortex, which receives tactile information from the surface of the body, the fingers are represented by more brain tissue than the arm, although they are smaller in surface area. Inouye also recognized that the occipital lobe contains neighbouring visual areas, with less precise topographic maps.

Credit for first describing the retinotopic map in the human brain is often mistakenly given to a British medical officer, Gordon Holmes, who performed a similar study during the First World War. It is unfortunate that war has often been the instrument of scientific progress. In his foreword, Inouye wrote: "The hardship and ferocity of the last war led me to publish these observations. The awfulness and horror of the experience, of which those who did not take part cannot have the slightest appreciation, at the same time raised the hope in me and in all other physicians that in the future, war may, if possible, be prevented." Tragically, as

Inouye wrote these words, the world stood on the brink of the bloodiest century in history. ■

Daniel L. Adams and Jonathan C. Horton are in the Department of Ophthalmology, University of California at San Francisco, 10 Kirkham Street, San Francisco, California 94143-0730, USA.

Supersymmetrical physics

The Quantum Theory of Fields: Volume III. Supersymmetry
by Steven Weinberg
Cambridge University Press: 2000. 419 pp.
£32.50, \$49.95

Hans Peter Nilles

The concept of quantum field theory, which arose in the 1930s from the combining of quantum theory with special relativity, enables us to describe all known phenomena of particle physics. The first two volumes of Steven Weinberg's book *The Quantum Theory of Fields* explain this brilliantly and have by now become classic textbooks.

As these two volumes showed, symmetries have had a crucial role throughout the development of quantum field theory. They allow a simpler formulation of these theories and lead to the appearance of conservation laws for physical quantities. The rotational invariance in three-dimensional space leads, for example, to the conservation of angular momentum, and the origin of conserved electric charge can be found in a symmetry that transforms the phase of the electron wavefunction. Sometimes new symmetries have been discovered by direct experimental observation; sometimes their presence has been imposed on physicists by demands for mathematical consistency and elegance. A cornerstone of the present-day 'standard model of particle physics' is the concept of gauge symmetry, which is used to describe electromagnetic interactions as well as the weak and strong nuclear forces.

Weinberg's third volume in the series, entitled *Supersymmetry*, describes a symmetry that was discovered as a mathematical curiosity some 30 years ago. Although there is no convincing evidence yet for its existence in nature, supersymmetry is one of the most discussed themes in modern particle physics. Intrigued by the uniqueness and theoretical beauty of this symmetry, practitioners in the field are confident that interest will continue in the coming years and even decades.

Supersymmetry relates two different classes of particles: bosons (which have

integer spin) and fermions (which have half-integer spin). In some sense, supersymmetry unifies objects as different as fire and water: that is, the quanta of radiation (bosons, such as the photon and the gluon) and the constituents of matter (fermions, such as the electron and quark). For a long time it was believed, as the consequence of a so-called 'no-go theorem', that such a boson-fermion symmetry could not exist. But as often happens, the loopholes in such 'theorems' led to surprises and the formulation of unique exceptions. Whereas conventional symmetries (such as gauge theories) can appear in infinitely many ways, only eight kinds of supersymmetry can be constructed. This 'uniqueness' is the consequence of the fact that supersymmetry enhances the symmetry of four-dimensional space-time in a non-trivial way, leading to the notion of 'superspace'. In fact, applying the gauge principle to the concept of supersymmetry leads to quantum field theories that automatically include Einstein gravity. This might be the right road to the long-sought unification of all fundamental interactions, including gravity.

Theories invoking supersymmetry have remarkable properties that are not shared by conventional quantum field theories: they have fewer divergences in perturbation theory (a type of theory used, for example, for calculating energy levels in atoms and molecules) and allow the formulation of 'non-renormalization theorems' for estimating quantities such as strength of interactions and particle masses. Divergences in quantum field theory, where some quantity tends towards infinity, make it sometimes impossible to compute physical quantities reliably. The non-renormalization theorems thus improve the power of calculability. The most symmetric of these supersymmetric theories can be shown to be finite. This makes supersymmetry a useful tool for finding exact solutions in quantum field theory and for developing mathematical methods that go beyond a description in terms of perturbation theory, with its infinite series of correction terms. Many of the most exciting developments of the past decades in mathematical physics, quantum field theory and string theory were obtained using the concepts of supersymmetry.

The presence of the non-renormalization theorems in supersymmetric quantum field theories lies at the origin of the conjecture that supersymmetry exists in nature and could be important in particle-physics phenomena at high energies and small distances. The current standard model of particle physics contains an intrinsic scale related to the strength of the weak nuclear force: $M_{\text{weak}} \approx 300 \text{ GeV}$, which determines the size of all known masses. This scale M_{weak} is much smaller than other scales in subnuclear

physics, most notably the Planck scale, $M_{\text{Planck}} \approx 10^{19} \text{ GeV}$, which governs the strength of gravitational interactions. In conventional quantum-field theories the small scale is usually destabilized by quantum corrections and driven towards the larger scales. The question, 'why is M_{weak} so small (and stable) compared with M_{Planck} ?' is known as the 'hierarchy problem' in particle physics.

Supersymmetry introduces an important ingredient into this discussion: some of the finiteness properties connected with the non-renormalization theorems can guarantee the stability of the weak scale. This then leads to a generalization of the standard model with the introduction of many new particles, the hypothetical 'supersymmetric partners' of known particles—gauginos, the fermionic partners of gauge bosons, for instance; or squarks and sleptons, bosonic partners of quarks and leptons. Such new particles should not be too far away from the fundamental scale $M_{\text{weak}} \approx 300 \text{ GeV}$, and could in principle be detected by high-energy experiments in operation or planned around the world. The lightest of these new hypothetical particles is expected to be stable and could be one of the particles that make up the dark matter in the Universe.

Unfortunately, up to now no direct experimental consequences of supersymmetry have been found. Experiments in the next decade will decide whether supersymmetry is myth or reality.

The third volume of *The Quantum Theory of Fields* is a self-contained introduction to the world of supersymmetry and supergravity. It will be useful both for experienced researchers in the field and for students who want to take the first steps towards learning about supersymmetry. Unlike other books in this field, it covers the wide spectrum of possible applications of supersymmetry in physics. Of course, the particular way in which supersymmetry might be realized in nature is not known, but we can hope that the not-so-distant future will reveal it. Such a discovery would be a triumph for theoretical physics. ■

Hans Peter Nilles is at the *Physikalisches Institut, Universität Bonn, Nussallee 12, D-53115 Bonn, Germany.*

More on particle physics Anomalies in Quantum Field Theory

by Reinhold A. Bertlmann
Oxford University Press, £29.95, \$50 (pbk)

New in paperback

Monsters of the Sea: The Truth About the Loch Ness Monster, the Giant Squid, Sea Serpents, Mermaids and other Fantastic Creatures of the Deep

by Richard Ellis
Globe Pequot, \$22.95

"Although it is not essential reading, every marine biological library and larger public library should have it. It will be useful also to those interested in the phenomenon of mass human self-deception, including psychologists and media moguls ... Ellis's book weeds out the myths, leaving an account of genuine monsters that, because they really exist, are more impressive than their fictional counterparts. Like all good science, it is undeniably fun." Paul F. S. Cornelius, *Nature* **375**, 746 (1995)

The Genetic Gods: Evolution and Belief in Human Affairs

by John C. Avise
Harvard University Press, \$17.95, £12.50

The Deep Hot Biosphere: The Myth of Fossil Fuels

by Thomas Gold
Springer, \$20, £15

"The book is clearly one man's view of major processes in the Earth sciences, and demonstrates that scientific debate is alive and well. Science is not just about facts and data, but is hypothesis-led and thrives on controversy —

and few people are more controversial than Thomas Gold." R. John Parkes, *Nature* **401**, 644 (1999)

The Advent of the Algorithm: The 300-year Journey from an Idea to the Computer

by David Berlinski
Harvest Books, \$14

Surfing Through Hyperspace: Understanding the Higher Universe in Six Easy Lessons

by Clifford A. Pickover
Oxford University Press, £8.99, \$14.95

Linnaeus: Nature and Nation

by Lisbet Koerner
Harvard University Press, \$19.95, £13.95
"A glance at Koerner's bibliography suggests that a substantial proportion of scholarly work on Linnaeus continues to be published in Swedish, a fact that corroborates one of her central assertions. If all politics is local, all science is too."
Harriet Ritvo, *Nature* **404**, 705–705 (2000)

A Mathematical Mystery Tour: Discovering the Truth and Beauty of the Cosmos

by A. K. Dewdney
Wiley, \$15.95, £11.50

At cross purposes

How do we cope with scientific terms that have two different definitions?

David M. Wilkinson

Technical terms abound in science; indeed, much of the science taught in schools seems more concerned with the memorizing of terms than with the understanding of concepts. A common defence of this complex scientific language is that it allows accurate communication among those who have mastered the appropriate vocabulary. Robin Dunbar, in *The Trouble with Science* (Faber, 1995), writes: "The need to have words that refer to specific phenomena is extremely important in science: failure to do so results in unnecessary talking at cross purposes, wastes people's time and holds up the advancement of science." However, the reality of the practice of science is more complex than suggested by such common-sense prescriptions.

A fascinating example is the term 'symbiosis', a biological term that also features in ordinary English. Even if its use in everyday English might be ambiguous, one would expect that there would be a single rigorous definition for its technical use. However, this expectation is wrong. From shortly after its introduction by Anton de Bary in the mid-nineteenth century, symbiosis (from the Greek: *syn* = together, *bio* = life) had two different scientific meanings. The first refers to a close physical association between two organisms, usually from different species; the second suggests a mutually beneficial relationship between the two organisms — 'mutualism' is another term used for the same purpose. Consider the roots of a plant: these can be infested by many species of fungi, some parasitic and some involved in mutually beneficial relationships with the plant. Both types are involved in a symbiosis according to the first definition, but only the beneficial fungi are symbiotic according to the second definition.

Symbiosis was coined by de Bary, a leading German plant pathologist, to describe close, long-term associations between different organisms. He was well known for his work on potato blight, so it's not surprising that he explicitly included parasitism as a type of symbiosis. However, by the early twentieth century both definitions were established in the scientific literature. For example, Eugene Warming, writing in 1909, followed de Bary's original definition, using 'mutualism' to describe beneficial relationships. Yet in a 1915 textbook, T. W. Woodhead defined symbiosis as "the union of two organisms whereby they mutually

benefit". Clearly, having two different definitions of the same term gives scope for much confusion. In a now-classic animal-ecology textbook of 1949, W. C. Allee and colleagues described both definitions but state a preference for the original one, writing "the broad meaning of symbiosis is the original one ... The term 'mutualism' in our usage corresponds exactly to the more limited concept of symbiosis that has been widely current."

Such attempts to standardize the definition of symbiosis have so far failed. I recently sampled 12 university-level textbooks of ecology and environmental science. Only three of them gave both definitions. The others gave only one meaning without mentioning the existence of an alternative usage; five followed the de Bary definition and four the mutualistic one. These confusions are also to be found among research biologists. I have several times been involved in confusing discussions before realizing that two different definitions of symbiosis were being used in the conversation. This confusion is mirrored in everyday English. Most general dictionaries I have examined stress cooperation and mutual benefit in their definitions, but I have also seen the close-association usage in some dictionaries.

An obvious question is, which is the better definition of symbiosis? The broader, living-together usage has the apparent advantage of historical precedence and it also appears to be the favourite one in the recent technical literature. It has the additional advantage of focusing attention on a fascinating evolutionary question: why some symbionts are parasites while others are mutualists, and why some organisms switch roles.

A wider question concerns the nature and use of scientific terminology. Does it matter that a term has two different definitions? Symbiosis is not the only example. 'Biosphere', for instance, is defined both as a location of life on Earth and as a system, rich in feedback loops, which links life with non-living aspects of its environment. It is obviously important that writers are clear how they are using such terms and that the authors of textbooks warn of the potential for confusion.

The simple taxonomist's approach of giving priority to first usage doesn't work for



The nitrogen-fixing *Rhizobium* (top) and the parasitic fungus that causes potato blight can both be classed as symbiotic, but only according to one of the two definitions of the term.

terms that apply to scientific ideas, because such terms need to change as our ideas evolve. Consider James Lovelock's concept of Gaia. This has developed since he first proposed it and there are currently several conflicting definitions of 'Gaia' in use reflecting the difficulties of understanding the complexities of the Earth system. With symbiosis, the textbook distinction between mutualism and parasitism is not always clear in nature, so it's perhaps unsurprising that the terminology mirrors these confusions.

David M. Wilkinson is at the School of Biological and Earth Sciences, Liverpool John Moores University, Liverpool L3 3AF, UK.

FURTHER READING

Douglas, A. E. *Symbiotic Interactions* (Oxford Univ. Press, Oxford, 1994).

Margulis, L. *The Symbiotic Planet* (Weidenfeld & Nicolson, London, 1998).

ANDREW SYRED/SPL DR JEREMY BURGESS/SPL

Climbing life's tree

Graham Budd

Fossils have always been a bother. Initially, natural philosophers were more impressed by their stony composition and where they were found than by what they looked like. Accordingly, they were compared to gemstones as often as to living organisms — perhaps not the best start for palaeobiology. Even when fossils were recognized as the remains of past life, no one knew how to classify them. Dinosaurs, ammonites and trilobites seemed to be quite like other reptiles, cephalopods and arthropods. But which ones were they like in particular? Conscientious palaeontologists strained sinews trying to force these groups to behave. Surely trilobites were a type of crustacean? Or did those antennae make them insects?

As these efforts at classification often failed, palaeontologists changed tack, creating countless high-level categories for fossils. At best, problematic groups were tagged as, for instance, 'annelid-like', given their own class or phylum, and cheerfully connected to the tree of life with dotted lines and question marks. This gave rise to the view that early evolution was different from 'standard' microevolution, with living groups of organisms suddenly appearing amid fireworks of excess 'body plans'. The most popular victim of this muddle has undoubtedly been the origin of animals in the 'Cambrian explosion'. Yet this amazing pattern — the inspiration for entire books devoted to analyses of its supporting mechanisms — is entirely the consequence of bad systematics.

Fossils become awkward if one imagines that our systematic groupings of extant biota somehow reflect the evolutionary process in its entirety — a task that they are ill-equipped to perform. Consider two closely related living groups, birds and crocodiles, neither of which gave rise to the other.

By definition, the most recent common ancestor of all living birds was a fully fledged bird, and the last common ancestor of all living crocodiles was, *mutatis mutandis*, a fully fledged crocodile. But what was the most recent common ancestor of both birds and crocodiles?

On the basis of the systematics of 'living forms', the possible answers seem to be bird, crocodile, or both. As the first two options are ruled out by definition (neither gave rise to the other), one would have to say that this animal fell at the base of both the bird and crocodile lineages. But from the definitions above, the two animals at the bases of these groups were already completely bird-like and crocodile-like, respectively. If they were identified as the same animal, it would have to have been grotesquely (if memorably) endowed with all the features of both. If your appetite for the bizarre is not sated by such a monster, consider the common ancestor of this grouping and snakes.

The ultimate view of evolution under this scheme would be that the ancestor of life somehow appeared with the features of all its descendants, and simply went on to abandon unnecessary lungs or fins as required. This preformationist model is patently absurd, but subconscious acceptance of its premises continues to have a surprisingly widespread and pernicious effect. Our perception of systematics exerts a profound influence on our view of evolution.

The division of fossil and living organisms into 'stem' and 'crown' groups, devised by the systematist Willi Hennig and the palaeontologist Dick Jefferies in the 1960s and 1970s, encapsulates the evolutionary process. Surprisingly, it has been applied very little. In this scheme, the most recent common ancestor of living birds and croco-

Stem groups

The division of fossil and living organisms into 'stem' and 'crown' groups encapsulates the evolutionary process, but has been applied very little.

diles represented not both, but neither of the living groups, and as the two lineages diverged from each other, their respective features accumulated in a step-by-step manner. The animals that filled the gap between the most recent common ancestor of the two living groups and that of only one of them form the latter's 'stem group'. Conversely, the most recent common ancestor of one of the living groups plus all of its descendants form the 'crown group'. The nice thing about stem groups for palaeontologists is that they must, by definition, comprise only fossil organisms (living forms are all in crown groups). Of course, stem groups themselves sit within larger crown groups.

If you are interested in how a living group evolved its distinctive features, and the principal transitions that led to it, you must look at the fossils of its stem group — the living forms are no help. Dinosaurs and trilobites can now find their rightful homes — in the stem groups of birds and chelicerate arthropods, respectively — without any strain; they also allow the reconstruction of ecological, functional and even developmental histories of the living groups in whose stem groups they are situated.

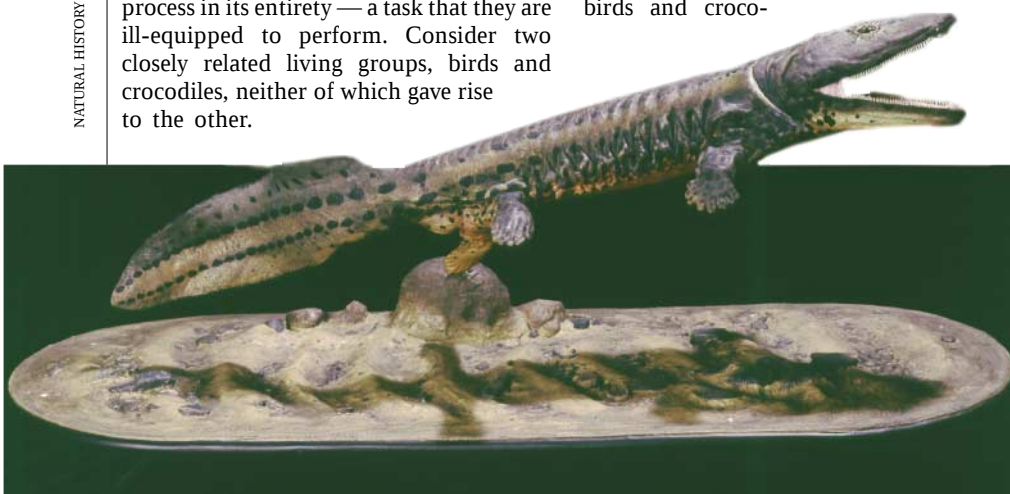
The low-tech study of morphology is not currently a popular discipline, but this does not render it unimportant. Morphology represents not just the product of the genes, but the mysterious interface between the genetic material and the environment — a prism through which the glare of selective pressure on the genome is refracted. If we wish, for example, to understand the evolution of the genome, we must also consider what morphology has been doing, probably in ways that we hardly suspect. If so, the secrets of the fossil record should be of interest in even the most up-to-date molecular laboratories. The classification of trilobites might yet find its place in the lexicon of the genomic revolution. ■

Graham Budd is in the Department of Earth Sciences, Uppsala University, Norbyvägen 22, Uppsala SE-752 36, Sweden.

FURTHER READING

Budd, G. E. & Jensen, S. *Biol. Rev.* **75**, 253–295 (2000).
Smith, A. *Systematics and the Fossil Record: Documenting Evolutionary Pathways* (Blackwell Scientific, Oxford, 1994).

NATURAL HISTORY MUSEUM, LONDON



Four-limbed forebear: the tetrapod *Acanthostega* is thought to be one of the earliest land vertebrates.

Flexible electronic futures

Robert J. Hamers

Microelectronic devices incorporating organic materials could find a host of applications. That prospect inches nearer with the development of a strategy for growing thin films of the organic semiconductor pentacene.

Virtually all of today's microelectronic devices are made from inorganic materials — silicon and gallium arsenide, for instance. But in the past few years there has been an explosion of interest in devices made from organic molecules for application in devices such as mechanically flexible computers and displays, and even truly 'molecular' computers in which individual molecules replace today's transistors^{1–3}. Organic semiconductors have been known for more than 20 years⁴, and last year's Nobel Prize in Chemistry⁵ was awarded for work in the area. But achieving electronic properties comparable to those of inorganic materials requires high-quality crystals^{6,7}, and these cannot easily be made in the necessary thin-film form by the vapour-deposition methods typically used in microelectronics fabrication. Thin films are needed because they are especially amenable to large-scale circuit fabrication.

On page 517 of this issue⁸ Meyer zu Heringdorf and co-workers report experiments in which they have imaged the

formation of thin films of pentacene, an important molecular semiconductor. Their experiments have produced a notable advance in making thin films of organic semiconductors with the required crystallinity, and led to a new understanding of the factors controlling the growth of these materials as thin films.

Meyer zu Heringdorf *et al.* used a technique known as photoelectron emission microscopy, or PEEM, to study film formation. In PEEM, the sample is illuminated with ultraviolet light, and photoelectrons that are ejected are imaged using a high-resolution electromagnetic lens system. PEEM provides sub-micrometre spatial resolution and can discriminate between different chemical species (such as silicon and a layer of organic molecules) if they emit photoelectrons at different rates. The imaging capability enables the authors to make movies of pentacene thin films forming on silicon surfaces.

Pentacene consists of five

fused aromatic — benzene-like — rings (Fig. 1). It is one of the most promising candidates for organic electronics, in part because of its chemical and thermal stability, and in part because its planar shape facilitates crystalline packing. By directly observing the surface while pentacene molecules were deposited, the authors were able to glean insight into the microscopic processes controlling crystal formation, and use that insight to achieve a technological advance.

When a simple inorganic material such as silicon or gallium arsenide is deposited onto a surface, the atoms first diffuse across the surface. When enough atoms pack together they form a stable 'nucleus', and further deposition creates 'islands' on the surface. Because organic molecules have more complex shapes and weaker forces binding them into crystals than do their inorganic counterparts, it has not been clear to what extent these classical models of growth can be applied to molecular materials. For example, pentacene (and probably most other molecular semiconductors) can form strong chemical bonds to the silicon surface, and even weak interactions can inhibit diffusion and modify the optimal orientation and packing of the molecules^{9,10} (Fig. 1a).

Using PEEM, Meyer zu Heringdorf *et al.* found that strongly bonded molecules inhibit nucleation and decrease the crystalline quality of the resulting pentacene film. Why? One likely reason is the close connection between chemical bonding and molecular shape. Whenever molecules break or form bonds, their shape changes. The changes in shape that occur when organic semiconductors (such as pentacene) form bonds to silicon are likely to be especially large. The formation of well-defined crystals requires molecules to nestle together like spoons in a drawer, so even small changes in shape prevent other molecules from packing together with them to form a crystal.

Ideally, film growth will be optimal when the strength of the interaction between the surface and the deposited molecules is strong enough to hold the molecules on the surface, but weak enough to allow them to diffuse and to rotate. To decrease the reactivity between the silicon surface and the pentacene molecules, Meyer zu Heringdorf and colleagues

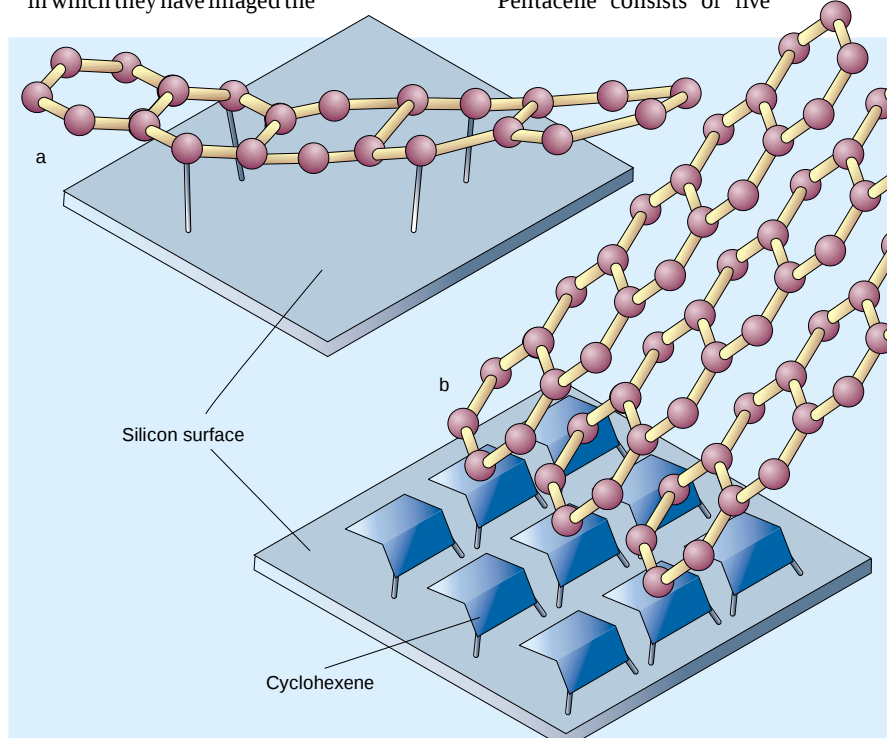


Figure 1 How to make thin-film crystals of pentacene, an organic semiconductor. a, A pentacene molecule that has attached directly to the silicon substrate forms strong bonds to the substrate and so is distorted, inhibiting crystal development. b, The approach used by Meyer zu Heringdorf and colleagues⁸. The silicon has been modified by cyclohexene molecules, which form two bonds with the silicon and coat it with an array of non-reactive C–H bonds. The pentacene molecules can then attach in ordered fashion to the cyclohexene, allowing stable crystal growth.

took advantage of a method for forming ordered organic cyclohexene monolayers on semiconductors¹¹; this method leaves the silicon surface coated with an ordered array of non-reactive C–H groups. The authors hypothesized that any interaction with the deposited pentacene molecules would be weak, and that this would lead to improved growth (Fig. 1b). Their hunch was right. Deposition of pentacene onto these organic-modified surfaces dramatically improves the crystallinity, leading to perfect thin-film crystals approaching 0.1 mm in size — much larger than the size of a single transistor. Similar results were obtained on oxidized silicon surfaces, provided that all of the chemically reactive ‘dangling bonds’ were eliminated. In essence, the new results show that the key to fabricating organic single crystals in thin-film form is to modify the substrate to reduce the bond strength to the organic molecules.

The technological significance of the new results is this: because microelectronic devices are only micrometres (or less) across, it should now be possible to make integrated circuits in which each transistor is made from a single crystal of pentacene, with corresponding improvements in electrical performance. But the results also have broader implications. The trend towards ‘molecular’ electronics is being played out on many length scales, from the micrometre scale of classical microelectronics using organic molecules, down to the use of single molecules³. There is tremendous potential for enhancing chemists’ abilities to design and fabricate ‘designer molecules’ with very specific types of electrical functionality. For instance, specific chemical groups might control the conductivity through a molecule, while other groups might be readily switched between oxidized and reduced states to make a kind of molecular switch^{1–3}.

The practical limitations in this enterprise may not lie in the materials themselves. Rather, they may be in understanding and controlling the properties at the interfaces between organic molecules and the inorganic materials (silicon, metals and so on) that connect them to the macroscopic world¹², and in learning how to coax molecules into assembling themselves into more complex structures. The new work⁸ extends the classical theories of nucleation and growth to more complex organic materials, thereby providing a conceptual link between the two. Given a similar understanding of the electronic properties of the materials and their interfaces, the dreams of cheap, flexible (and perhaps even biodegradable) microelectronics based on organic molecules may soon become a reality. ■

Robert J. Hamers is in the Department of Chemistry, University of Wisconsin–Madison, 1101 University Avenue, Madison, Wisconsin 53706, USA. e-mail: rjhamers@facstaff.wisc.edu

1. Aviram, A. & Ratner, M. *Chem. Phys. Lett.* **29**, 277–283 (1974).
2. Collier, C. P. *et al. Science* **287**, 391–394 (1999).
3. Joachim, C., Gimzewski, J. K. & Aviram, A. *Nature* **408**, 541–548 (2000).
4. Chiang, C. K. *et al. Phys. Rev. Lett.* **39**, 1098–1101 (1977).
5. <http://www.nobel.se/chemistry/laureates/2000>
6. Crone, B. *et al. Nature* **40**, 521–523 (2000).
7. Schön, J. H., Berg, S., Kloc, Ch. & Batlogg, B. *Science* **287**, 1022–1023 (2000).
8. Meyer zu Heringdorf, F.-J., Reuter, M. C. & Tromp, R. M. *Nature* **412**, 517–520 (2001).
9. Dimitrakopoulos, C., Brown, C. & Pomp, A. *J. Appl. Phys.* **80**, 2501–2508 (1996).
10. Kasaya, M., Tabata, H. & Kawai, T. *Surf. Sci.* **400**, 367–374 (1988).
11. Hamers, R. J. *et al. Acc. Chem. Res.* **33**, 617–624 (2000).
12. Pantelides, S. T., DiVentra, M. & Lang, N. D. *Physica B* **296**, 72–77 (2001).

Demography

Uncertain population forecasts

Nico Keilman

Traditional population forecasts made by statistical agencies do not quantify uncertainty. But demographers and statisticians have developed methods to calculate probabilistic forecasts.

The demographic future of any human population is uncertain, but some of the many possible trajectories are more probable than others. So attempts to forecast demographic aspects of a population, such as its size by a given year, should include two elements: a range of possible outcomes, and a probability attached to that range. Together, these elements constitute ‘prediction interval’ for the population variable concerned. There is a clear trade-off between greater certainty (higher odds) and better precision (narrower intervals). For instance, on page 543 of this issue, Lutz, Sanderson and Scherbov¹ estimate that the odds are 4 to 1 (an 80% chance) that the world’s population, now at 6.1 billion, will be between 5.6 billion and 12.1 billion in the year 2100. Odds of 19 to 1 (a 95% chance) result in a wider interval: 4.3 billion to 14.4 billion.

Demographers have become increasingly concerned about the accuracy of their forecasts, in part because the rapid fall in fertility in Western countries in the 1970s came as a surprise. Forecasts made in those years predicted birth rates that were up to 80% too high and too many young children. The rapid reduction in mortality after the Second World War was also not foreseen; life-expectancy forecasts were too low by 1–2 years; and the predicted number of elderly, particularly the oldest people, was far too low^{2,3}.

Those who use forecasts should be informed about the accuracy of historical predictions. But even more important is the expected accuracy of the current forecast. Statistical agencies traditionally deal with the uncertainty of forecasting population variables by producing two or more predictions of fertility or mortality (or both), and then calculating a range of predictions. For instance, Statistics Norway expects the number of children aged 6–12 in Norway in 2010 to be between 401,000 and 436,000, depending on whether fertility is low or

high — that is, on whether women have an average of 1.5 or 2.1 children, respectively, in 2010. The agency attaches no probability to this interval. Yet those who are planning provisions for education need to know whether the likelihood of this scenario is roughly 30%, 60% or even 90%.

So, during the 1990s, demographers and statisticians developed methods for making probabilistic population forecasts, the aim of which is to calculate prediction intervals for every variable of interest. Examples include population forecasts for the United States, Austria, Germany, Finland and the Netherlands; these forecasts comprised prediction intervals for variables such as age structure, average number of children per woman and immigration flow^{4–8}. Lutz *et al.*¹ present probabilistic population forecasts for the whole world, and for the 13 large regions that they divide the world into (see Table 1 on page 544). Their most important message is that there is an estimated 85% chance that the world’s population will stop growing before 2100.

Yet probability statements of this kind — which give the expected accuracy of a population forecast — must be interpreted with care, because they vary depending on the statistical method used and its assumptions. For instance, Lutz *et al.*¹ estimate that there is a 95% chance that the world population will reach 6.6 billion to 11.4 billion in 2050 (see Fig. 2 on page 544). But the US National Research Council (NRC) uses a different method and has calculated that there is a 95% chance that the 2050 world population will be between 7.9 billion and 10.9 billion⁹.

There are three main methods of probabilistic forecasting: time-series extrapolation; expert judgement; and extrapolation of historical forecast errors. Time-series methods rely on statistical models that are fitted to historical data. These methods, however, seldom give an accurate description of the past. If many of the historical facts remain

unexplained, time-series methods result in excessively wide prediction intervals when used for long-term forecasting.

Judgemental methods can be used to correct or constrain such broad prediction intervals. Expert judgement is also used when expected values and corresponding prediction intervals are hard to obtain by formal methods. In demographic forecasting, the expert-judgement method has been pioneered by Lutz and colleagues^{5,6,10}. A group of experts is asked to indicate the probability that a key parameter (such as the average number of children per woman, or life expectancy) in some future year falls within a certain prespecified range. A weakness of this approach is that experts, often being unduly confident, tend to give over-optimistically narrow prediction intervals¹¹. When the forecasts are later compared with actual data, the intervals turn out to fit the observed trends much less frequently than the probabilities suggested.

Finally, empirical errors observed for past forecasts may be extrapolated to predict the expected errors for the current forecast. For the NRC report⁹, this method was included in the statistical model to calculate the uncertainty of United Nations population forecasts for every country. A problem here is that forecasts prepared in the 1960s or earlier were poorly documented, so data on historical errors do not stretch back as far as one would like.

Elements of the three methods are often used in combination. For instance, time-series methods involve some degree of subjectivity, perhaps in choosing the extrapolation model or the length of the historical data series. These decisions may strongly influence the prediction intervals. And the intervals, whether obtained by time-series methods or expert opinion, are frequently checked against historical error patterns^{4,7}. Lutz *et al.*¹ combined expert judgement with time-series models.

Irrespective of the method used, probabilistic forecasts of the youngest and oldest age groups show the most uncertainty, because fertility and mortality are hard to predict. Short-term forecasts of the demography of developing countries are also subject to great uncertainty, because of the poor quality of their demographic data. Finally, prediction intervals are always narrower when variables are aggregated (for example, when the populations of all countries are added together) rather than looked at individually, because the errors tend to cancel each other out. These kinds of uncertainty assessments are crucial, and statistical agencies would do a great service to users of forecasts if they would adopt probabilistic methods rather than more traditional methods that do not take uncertainty into account. Yet those making use of population forecasts should also keep in mind that

probabilistic statements themselves are uncertain to some extent.

Nico Keilman is in the Department of Economics, University of Oslo, Blindern, N-0317 Oslo, Norway.
e-mail: nico.keilman@econ.uio.no

1. Lutz, W., Sanderson, W. & Scherbov, S. *Nature* **412**, 543–546 (2001).
2. Horiuchi, S. *Nature* **405**, 744–745 (2000).
3. Keilman, N. J. *Official Stat.* **13**, 245–277 (1997).
4. Lee, R. & Tuljapourkar, S. *J. Am. Stat. Assoc.* **89**, 1175–1189 (1994).

5. Lutz, W. & Scherbov, S. *Eur. J. Pop.* **14**, 1–17 (1998).
6. Lutz, W. & Scherbov, S. *Z. Bevölkerungswissenschaft* **23**, 83–109 (1998).
7. Alho, J. *A Stochastic Forecast of the Population of Finland Reviews 1998/4* (Statistics Finland, Helsinki, 1998).
8. De Beer, J. & Alders, M. Working Paper no. 45 <http://www.unecce.org/stats/documents/1999.05.projections.html>
9. National Research Council *Beyond Six Billion: Forecasting the World's Population* (Nat. Acad. Press, Washington DC, 2000).
10. Hanika, A., Lutz, W. & Scherbov, S. *Statist. Nachrichten* **12**, 984–988 (1997).
11. Armstrong, J. *From Crystal Ball to Computer* 2nd edn (Wiley, New York, 1985).

Nanotechnology

Boning up on biology

T. Andrew Taton

Attempts to tailor nanometre-scale objects to mimic and interact with natural materials raise the question of how to predict the biological response to these tiny creations.

The science of manipulating matter on the nanometre scale — nanofabrication — has provided researchers with a remarkably diverse set of tools for probing the behaviour of biological structures. At a recent meeting* on the “Biological Applications of Nanotechnology”, scientists at the

*American Chemical Society ProSpectives Conference, Berkeley, California, USA, 3–6 June 2001.

forefront of this research showed how these tools are already being used to sequence genomes, diagnose disease and watch single biomolecules perform their natural functions in real time. But there were also a few hints of a fresh theme emerging from all this work: the design of artificial nanostructures that can interact with and replace natural, biological materials. Studies that aim at

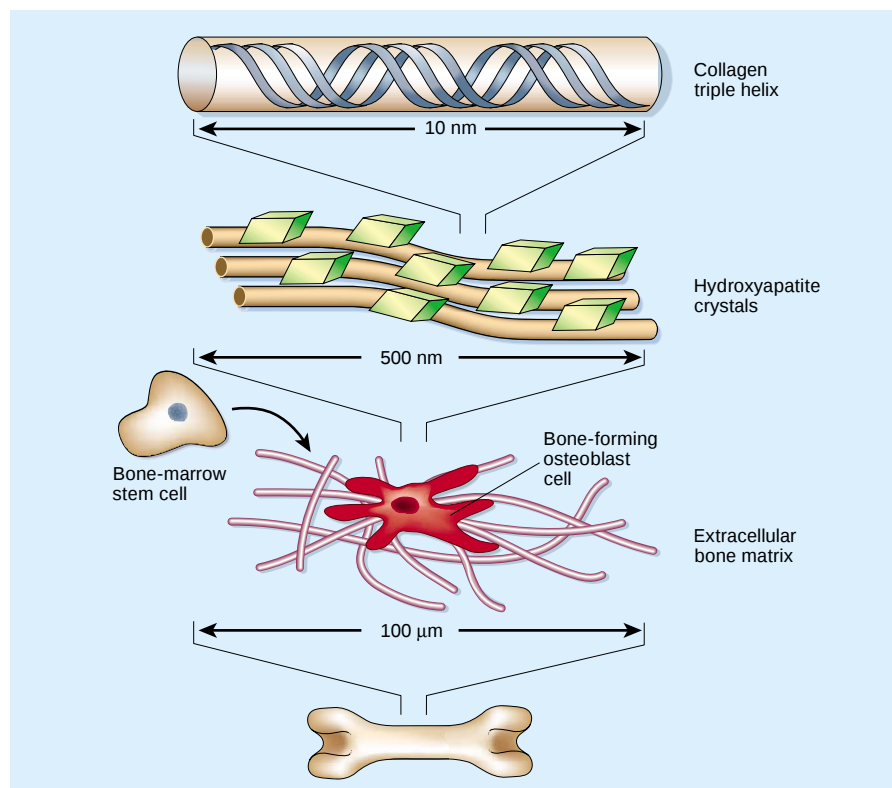


Figure 1 The many scales of organization in natural bone. Collagen triple helices spontaneously form nanoscale bundles of protein, which act as a template for the crystallization of hydroxyapatite nanocrystals. The collagen matrix is also recognized by undifferentiated bone-marrow stem cells, which become bone-forming osteoblasts after signals from bone-specific proteins in the matrix. To duplicate the extraordinary durability and strength of natural bone, researchers must design synthetic materials that match these structural levels of organization from the nanoscale upwards. (Note that this figure shows the dimensional, not sequential, principles of bone construction.)



100 YEARS AGO

Boomerangs may be studied for their anthropological interest as examples of primitive art, or for the manner in which they illustrate dynamical principles. But there is extraordinary fascination in making and throwing them, and in watching the remarkable and always graceful curves described in their flight; accordingly my chief object in the following paper has been to diminish the practical difficulties of the subject by giving some of the results of ten years' experimental acquaintance with it... It is rather difficult to give sufficient spin to keep the motion stable through a long flight, and I have found it advantageous to wind round the wood about 60 grammes weight of copper wire in three equal portions, of which one is near the middle and one near each end. This materially increases the moment of inertia about the centre of gravity without interfering seriously with other details. I have thrown a loaded boomerang of this type 167 metres, and my range with a spherical ball of half the weight is only 63 metres.

From *Nature* 1 August 1901.

50 YEARS AGO

The removal of the nucleus causes a slow decrease of the ribonucleic acid content of the cytoplasm of *Amoeba proteus*. This indicates that the ribonucleo-protein particles (microsomes) subsist in the cytoplasm only in the presence of the nucleus. However, this nuclear control is not immediate, for the ribonucleic acid content of non-nucleated fragments drops significantly only after the fourth day following enucleation. It appeared of interest to find out whether the oxygen uptake is under nuclear control; a rapid decrease of respiration after enucleation, as stated by Clark, might mean that respiratory enzymes, known to be predominantly bound to mitochondria, are dependent on the nucleus for the maintenance of their activity... One is therefore led to believe that enucleation leads to a break in the normal coupling between oxidations and phosphorylations: the non-nucleated amoeba would thus behave like cells treated with dinitrophenol, which interrupts synthesis by blocking the coupling between oxidation and phosphorylation. Work is now in progress to find out if the nucleus, and in particular the nucleolus, takes any part in the synthesis of the coenzymes necessary for this coupling.

From *Nature* 4 August 1951.

replicating the complex nanoscale structure of cell membranes, tissue and bone are leading to promising materials with potential uses as implants and therapies. At the same time, the work has the potential to illustrate how cells interact with nanometre-sized objects in their own world.

Bone is one example of a natural material whose properties depend intimately on its nanoscale structure. Bone is an inorganic–bioorganic composite material consisting mainly of collagen proteins and hydroxyapatite (a crystalline form of calcium phosphate). Collagen spontaneously forms fibrils of aligned protein helices¹, on which tiny hydroxyapatite crystals (10–50 nanometres in length) can grow² (Fig. 1). Both the size and the orientation of the crystals are dictated specifically by the collagen template, and the precise structural relationship between the collagen and hydroxyapatite is critical to bone's resilience and strength.

A group at Northwestern University has been trying to design self-assembling, synthetic substitutes for collagen, which can also act as templates for hydroxyapatite crystallization (S. Stupp, Northwestern Univ., Evanston, Illinois). Stupp demonstrated that amphiphilic molecules — bearing a long hydrophobic alkyl group on one end and a hydrophilic peptide on the other — spontaneously assemble into cylindrical structures that resemble collagen fibrils. More importantly, these cylinders guide the formation of hydroxyapatite crystallites with orientations and sizes similar to those in natural bone. It remains to be seen whether Stupp's composite materials replicate the mechanical properties of bone, and whether the nanoscale order in these materials translates into macroscale toughness in a prosthetic bone replacement. Nevertheless, the work represents a step towards synthetically matching the complex, hierarchical structure of bone.

Stupp's work also raises a question in the context of biologically orientated nanotechnology research. Can we synthesize materials that not only replicate the properties found in nature but also prompt biological systems to build on these materials? In natural bone formation, for example, bone-specific proteins^{3,4} and possibly some types of collagen⁵ have peptide sequences that signal bone-marrow stem cells to differentiate into bone-forming osteoblast cells (Fig. 1). The osteoblasts then supply the calcium required to form the crystalline hydroxyapatite. The initial results reported by Stupp were obtained with amphiphilic molecules terminated by Arg-Gly-Asp (RGD), a peptide sequence known to be a general promoter of cell adhesion and growth^{6,7}. However, these molecules cannot promote the differentiation of marrow stem cells.

The next challenge in this research will be to synthesize amphiphilic assemblies bearing other peptide sequences that specifically

bind marrow cells and trigger osteoblast differentiation. Once again, the nanoscale organization of these signalling sequences will be crucial to this effort; it has already been shown that the presence of the sequences alone is not enough to promote osteoblast differentiation or bone formation⁸. The technological goal of this research would be to replace solid bone grafts with an organic scaffold that would foster the growth of new, natural bone. In theory, an appropriately designed material could achieve this only by organizing the inorganic component of bone and by communicating specific messages to bone-forming cells.

Unfortunately, several other speakers at the conference stressed just how difficult it can be to design nanostructured materials to interact specifically with cell surfaces. For example, merely attaching peptides to nanoscale materials does not completely dictate their interactions with cellular membrane receptors; other properties, such as general biocompatibility, solubility and nonspecific interactions with cell membranes, have to be correctly engineered as well. These issues have made it particularly challenging to coerce cells to specifically recognize and ingest inorganic nanoparticles known as quantum dots (M. Bruchez, Quantum Dot Corp., Hayward, California; W. Parak and A. Alivisatos, Univ. California, Berkeley).

In the world of organic dendrimers — highly branched polymers that form spherical shapes in solution — even small variations in the size of dendrimer particles, over the range 1–10 nm, can have significant effects on the ability of these materials to carry molecules across cell membranes (D. Tomalia, Univ. Michigan). Because cells interact with many nanoscale objects in their own world (such as vesicles and viruses), attempts to engineer unnatural nanostructures such as quantum dots and dendrimers promise to reveal more about why natural nanostructures have the properties they do. Admittedly, there is still some way to go before nanotechnology and biology march hand in hand. Nonetheless, the science of synthesizing nanostructured materials that specifically recognize biological materials is advancing quickly, and points towards a future when nature will welcome our nanoscale inventions. ■

T. Andrew Taton is in the Department of Chemistry, University of Minnesota, Minneapolis, Minnesota 55455, USA.

e-mail: taton@chem.umn.edu

- Holmgren, S. K., Taylor, K. M., Bretscher, L. E. & Raines, R. T. *Nature* **392**, 666–667 (1998).
- Weiner, S. & Wagner, H. D. *Annu. Rev. Mat. Sci.* **28**, 271–298 (1998).
- Niklason, L. E. *Nature Biotechnol.* **18**, 929–930 (2000).
- Bruder, S. P., Fink, D. J. & Caplan, A. I. *J. Cell. Biochem.* **56**, 283–294 (1994).
- Mizuno, M., Fujisawa, R. & Kuboki, Y. *J. Cell Physiol.* **184**, 207–213 (2000).
- Pierschbacher, M. D. & Ruoslahti, E. *Nature* **309**, 30–33 (1984).
- Roberts, C. et al. *J. Am. Chem. Soc.* **120**, 6548–6555 (1998).
- Staat, W. D. et al. *J. Biol. Chem.* **266**, 7363–7367 (1991).

Gene regulation

Cycling silence

Leonie Ringrose and Renato Paro

The cell-division cycle involves careful timing: cell-cycle genes must be switched on only when needed. The retinoblastoma protein controls one set of these genes, and one way it does this is most unusual.

In the desert, vast tracts of land are lifeless, interrupted only by the occasional oasis. In our cells, part of the genome seems similarly devoid of life. Apart from a few oases of activity, most of the genetic information in this part is silenced—it is never switched on. This region, where the DNA is packaged into a closed, inactive structure, is called heterochromatin¹. This type of global silencing is thought to be rather unchanging, and at first glance seems inappropriate to the exquisitely timed control of the cell-cycle genes, which guarantee the smooth progression of a cell through the steps of DNA replication and cell division. But, on page 561 of this issue, Nielsen and colleagues² show that the two human proteins that are required for the formation of heterochromatin are also involved in the dynamically regulated repression of some cell-cycle genes. These proteins are brought into play by the retinoblastoma protein (Rb), known for its involvement in cell proliferation.

Heterochromatin is a special state of chromatin. In chromatin, DNA is wrapped around histone proteins, forming nucleosomes. The flexible tails of the histones protrude from the globular, DNA-wrapped nucleosome core. Specific chemical modifications of these tails, such as the addition or removal of phosphate, acetyl or methyl groups, can profoundly affect the architecture of the whole chromatin package, enabling it to open up or close down as crucial steps in regulating gene activity³.

In heterochromatin, an architecture that represses gene activity is ensured by a part-

nership of two proteins: one marks histone tails, and the other recognizes the mark. The marking protein, SUV39H1, methylates lysine residue 9 on the tail of histone protein H3 (Fig. 1a)⁴. Heterochromatin protein 1 (HP1) recognizes this mark and binds to it (Fig. 1b). After many iterations of this process, the result is a stable structure that extends over hundreds of thousands of bases and prevents gene activation^{5,6}. Remarkably, genes that are experimentally placed near to heterochromatin can also become silenced, regardless of their identity. This promiscuous silencing is caused by spreading of the heterochromatin structure.

Given the indiscriminate, static nature of heterochromatic silencing, Nielsen *et al.*'s discovery²—that SUV39H1 and HP1 are also involved in the dynamic regulation of specific cell-cycle genes by the Rb protein—is surprising. Rb acts as a gatekeeper at the transition from the G1 arrest phase of the cell cycle into the DNA-replication (S) phase (Fig. 2)⁷. It keeps cells in the G1 phase by silencing a fleet of genes needed for entry into the S phase. Accurate repression of S-phase entry is vital to correct proliferation, and is lost if Rb-mediated silencing is disrupted, as it is in many cancers.

One way in which Rb silences S-phase genes is by blocking the E2F protein, a gene-expression activator that binds to the control regions (promoters) of cell-cycle genes. Rb binds to E2F at promoters, blocking its activity. Rb also recruits other proteins to help it, several of which modify chromatin architecture. Examples include histone

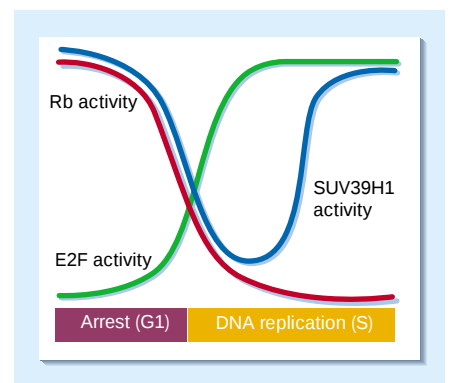


Figure 2 Silencing during the cell cycle. The transition from the G1 to the S phase of the cell cycle is a key checkpoint in cell proliferation. During the G1 phase, the retinoblastoma protein (Rb) blocks the transition by countering the activity of the E2F protein, silencing the genes necessary for entry into the S phase. At the end of G1, Rb is inactivated. The methylation activity of SUV39H1 during the cell cycle is also shown. This protein is inactivated at about the same time as Rb, but is reactivated later in S phase.

deacetylases, and other protein complexes that can move nucleosomes⁸. So Nielsen *et al.*² wondered, reasonably, whether Rb might enlist the services of yet more chromatin-modifying machinery, such as HP1 and SUV39H1.

The Rb protein works directly at the control regions of cell-cycle genes, none of which are found in heterochromatin. So Rb must physically bring HP1 and SUV39H1 to these genes if it is to make use of their silencing properties. Nielsen *et al.* first show that HP1, Rb and SUV39H1 can indeed stick together in living cells. But does this tripartite complex actually do anything to Rb's target genes? To find out, the authors compared two genes that are regulated by Rb or its relatives with genes that are not. They discovered that only the Rb-regulated genes (cyclin A and cyclin E) depend on the SUV39H1 protein for their repression. So the idea that Rb recruits SUV39H1 and HP1 to silence target genes gathered momentum.

To be sure that SUV39H1 and HP1 repress Rb's target genes in the same way that they repress genes in heterochromatin (Fig. 1a, b), Nielsen *et al.* went one step further. They looked for histone methylation at the cyclin E promoter, and found that Rb-dependent silencing is indeed accompanied by the methylation of histone H3 and the binding of HP1. The biggest surprise was that these events occurred at one nucleosome surrounding the promoter, but did not spread to the next nucleosome along. This tight localization is in stark contrast to the apparently promiscuous spreading of methylation and bound HP1 in heterochromatin.

This work raises several questions. From

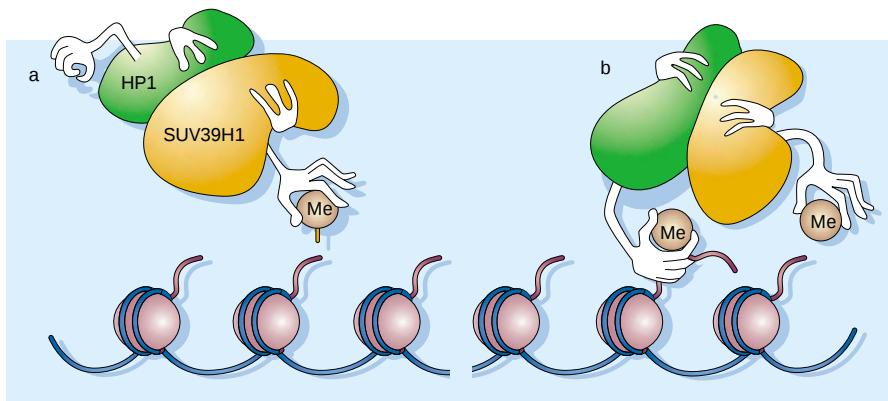


Figure 1 Silencing in heterochromatin, a compact form of DNA, by the HP1 and SUV39H1 proteins. DNA (blue) is packaged into nucleosomes by histone proteins (mauve circles). a, Protruding tails of histone H3 proteins are methylated by SUV39H1. b, The HP1 protein binds to the methylated tail and represses gene expression. This heterochromatin structure may spread if SUV39H1 methylates the next nucleosome along, which can in turn be bound by other HP1–SUV39H1 pairs.

the point of view of Rb-mediated gene silencing, one wonders why so many different mechanisms for changing chromatin architecture are involved. The presence of both histone deacetylases and the HP1 and SUV39H1 proteins at the cyclin E promoter might seem excessive. However, as Nielsen *et al.* point out², the deacetylase may be needed to remove an acetyl group from lysine 9 of histone H3, so that the lysine can be methylated by SUV39H1. This may be a general principle, so that we would expect a histone deacetylase to be needed wherever there is a methylase. Alternatively, this combination might be unique to cyclin E, with different mechanisms operating at other targets of Rb. If SUV39H1 does have a general role in Rb-mediated silencing, then we would expect any disruption of this protein to drive a cell to proliferate excessively, perhaps leading to cancer. Indeed, the loss of SUV39H1 or its activity does lead to hyperproliferation and other cell-division defects^{4,9}.

From the point of view of heterochromatin, the most interesting question is: why does histone methylation and gene silencing, as controlled by the HP1–SUV39H1 partnership, spread over hundreds of kilobases in heterochromatin, but does not even travel from one nucleosome to the next at the cyclin E promoter? One answer might lie in the way in which SUV39H1 is regulated during the cell cycle (Fig. 2). Like Rb, SUV39H1 is active during the G1 phase and is inactivated at the G1-to-S transition by phosphorylation. But unlike Rb, the inactivation of SUV39H1 is temporary, being lifted by late S phase⁹. Heterochromatin is replicated late in S phase, whereas other genes are replicated earlier. Perhaps newly replicated heterochromatin receives a heavy dose of methylation and HP1 in late S phase, whereas the cyclin E promoter (which is replicated when SUV39H1 is inhibited) escapes with just a little.

Finally, the fact that this mode of silencing is not confined to heterochromatin raises the possibility that the inner life of heterochromatin itself is more dynamic than we thought. Perhaps there is life in this desert after all.

Leonie Ringrose and Renato Paro are at the Center of Molecular Biology (ZMBH), University of Heidelberg, Im Neuenheimer Feld 282, 69120 Heidelberg, Germany.

e-mail: paro@sun0.urz.uni-heidelberg.de

- Wallrath, L. L. *Curr. Opin. Genet. Dev.* **8**, 147–153 (1998).
- Nielsen, S. J. *et al. Nature* **412**, 561–565 (2001).
- Strahl, B. D. & Allis, C. D. *Nature* **403**, 41–45 (2000).
- Rea, S. *et al. Nature* **406**, 593–599 (2000).
- Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. *Nature* **410**, 116–120 (2001).
- Bannister, A. J. *et al. Nature* **410**, 120–124 (2001).
- Brehm, A. & Kouzarides, T. *Trends Biochem. Sci.* **24**, 142–145 (1999).
- Zhang, H. S. *et al. Cell* **101**, 79–89 (2000).
- Firestein, R., Cui, X., Huie, P. & Cleary, M. L. *Mol. Cell. Biol.* **20**, 4900–4909 (2000).

Superconductivity

Is kinky conventional?

Philip B. Allen

An explanation for superconductivity in high-temperature, or 'unconventional', materials remains elusive. New experiments may indicate a surprising similarity to their low-temperature counterparts.

Copper oxides are still by a factor of three the highest-temperature superconductors around. Superconductivity requires the binding of electrons in pairs, but the binding force in these 'unconventional' materials remains unknown. An analysis of photoelectron spectra, described by Lanzara *et al.* on page 510 of this issue¹, reveals a 'kink' in the spectrum of low-energy electrons in copper oxides. The authors interpret this kink as the signature of interactions between electrons and vibrations of the material lattice, known as phonons. According to the accepted theory, the same interactions control conventional superconductors (which operate at much lower temperatures). If Lanzara and colleagues' interpretation is correct, it throws much previous thinking into doubt and indicates that the electron–phonon interaction may be relevant to unconventional superconductivity.

First, however, some background. Beginning in the 1930s, Landau argued that continuous phase transitions have common features, most simply exhibited in the magnetic phase transition of a ferromagnet such as iron. When temperature T is reduced to a transition temperature T_c , a new field ϕ (magnetization in this case) appears. For ϕ to become stabilized and permanent at

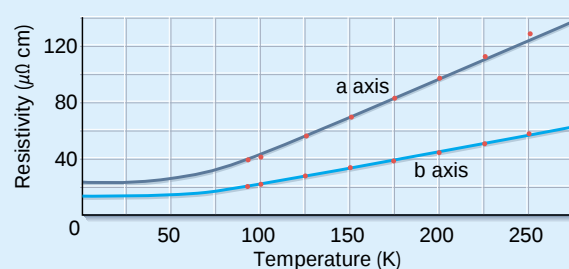
low temperatures, there has to be an interaction that causes $\phi(r)$ at one place (r) to stabilize $\phi(r')$ at nearby places (r'). So, for superconductivity — where T_c is the temperature at which the material loses resistance to current flow — the first aim for a theory was to identify ϕ and the interaction involved.

Following the discovery of superconductivity in 1911, both field and interaction remained a mystery for almost half a century. Then, in 1957, the Bardeen, Cooper and Schrieffer (BCS) theory introduced a new form of quantum coherence into physics. It showed that the field ϕ is a quantum 'pair field', or 'pair wavefunction', and that the interaction had to be an attraction between electrons.

Normally one imagines that because electrons have negative charge, they must repel each other. So how, in a superconducting state, do they attract each other and pair up? Building on earlier theories and experiments, Bardeen, Cooper and Schrieffer identified the source of the attraction as an 'indirect' interaction, in which one electron alters a vibration of the material lattice (a phonon) and the other electron feels the alteration. Like two ships floating close together at sea and modifying the surrounding waves, the

Box 1 Resistivity in high- T_c metals

The following, necessarily technical, account addresses a widespread misconception about resistivity in high- T_c metals. The temperature dependence of resistivity (ρ) of many metals is caused by electron–phonon interactions, as described by the Bloch–Grüneisen formula. It is not widely appreciated that the resistivity of the high- T_c superconductors is actually fairly conventional in terms of temperature dependence. The figure¹⁹ here shows the measured resistivity²⁰ in two crystal directions (a and b) for a single crystal of $\text{YBa}_2\text{Cu}_3\text{O}_7$ and the corresponding Bloch–Grüneisen curves. $\text{YBa}_2\text{Cu}_3\text{O}_7$ is superconducting



below a temperature of 92 K. The electron–phonon parameter $\lambda = 1$ is adjusted to give the best fit. The difference between the slope of ρ versus T in the two crystal directions is accounted for by the predicted^{21,22} inverse effective mass tensor $(n/m)_{\text{eff}}$. Perhaps this good fit is only an accident.

However, it should be acknowledged that resistivity can be accommodated remarkably well by electron–phonon theory, so reason (3) cited in the main text can hardly be used to exclude the electron–phonon mechanism for superconductivity of copper oxides.

P. B. A.

electrons attract each other. The interaction causing conventional superconductivity is, then, the electron–phonon interaction.

From the beginning, it was apparent that BCS theory can accommodate other types of attractive interactions, and that the pair field can belong to a variety of symmetry classes. For example, the *s*-wave symmetry of electron pairing in conventional superconductors might be replaced by more complicated symmetries such as *p* or *d* wave. A search for exotic forms of superconductivity has filled many volumes and covered all known classes of inorganic and organic solids. But only relatively recently has it yielded firm evidence for pair fields more complex than the simplest *s*-wave form. Since 1993 we have known that the ϕ of the high- T_c copper oxide superconductors has *d*-wave symmetry². However, the interaction responsible remains a matter of debate: neither theorists nor experimentalists have been able to produce a convincing picture of what might be happening.

Lanzara *et al.*¹ carried out experiments in three different types of copper oxide superconductor. They looked at energy (E_k) as a function of momentum (k) for the least energetic photoholes created by tuned photons from a synchrotron light source. These holes are empty states formerly occupied by ejected electrons. They reside in the highest occupied energy levels, the ones directly responsible for superconductivity. Technical advances mean that fine details of the spectrum can be resolved, including a characteristic energy shift caused by coupling of holes to phonons. This shift results in a ‘kink’ in E_k versus k (see Fig. 1 of the paper on page 511). Such a kink was previously seen in one copper oxide superconductor³, but Lanzara *et al.* find that it is ubiquitous in high- T_c metals. The effect has been studied in conventional metals^{4–6} and is well understood. But until now there has been little evidence for it in the copper oxides, and the phonon interpretation has not previously been invoked.

This is why Lanzara and colleagues’ interpretation will be controversial: the majority view is that the electron–phonon interaction of conventional superconductors is, at most, peripherally relevant to high- T_c superconductivity. Various reasons are given. (1) T_c is too high for a phonon mechanism. (2) Pair fields with *d*-wave symmetry are not favoured by phonons. (3) Properties such as resistivity are not explained by electron–phonon interactions. (4) All copper oxide superconductors show fluctuating antiferromagnetic order, which strongly suggests that unconventional physics is involved. (5) The layered crystal structures of the copper oxides suggest that special two-dimensional physics is important, whereas phonons in two dimensions are probably no different from those in three.

None of these arguments is watertight. The evidence for (4), which casts the greatest doubt on electron–phonon mechanisms, and (5) is still largely circumstantial.

In its simplest form, (3) is plain wrong (see Box 1). The first two can be countered in various ways. First, conventional theory is less restrictive in T_c values than originally thought⁷. Second, to get *d*-wave pairing one needs anisotropic coupling, for example, enhanced interactions for electrons that scatter with either no momentum transfer^{8–10} or complete momentum exchange¹¹. Or one could argue that conventional theory needs to be enlarged because of the low carrier density of copper oxides^{12,13}, possibly leading to isolated (bipolaronic^{14–16}) bound pairs of electrons, rather than the overlapping, delocalized pairs of conventional superconductivity.

On the one hand, then, the reasons for dismissing Lanzara and colleagues’ conclusions may not stand close scrutiny. On the other, it is hard to account for the previous failure of electron spectroscopists to identify electron–phonon effects in high- T_c materials. Two groups^{17,18} have made detailed theoretical calculations using density functional theory to explain the resonant frequencies of lattice vibrations. These theories produce numerical values of the electron–phonon coupling strength that are a by-product of careful and otherwise highly successful calculations, and cannot be dismissed lightly. The computed values are much too large to be invisible experimentally, and yet much too small to account for the high- T_c *d*-wave order parameter.

It will be interesting to follow the arguments that ensue after publication of this paper¹. They may not lead to identification of the interaction mechanism of high- T_c superconductivity, but they should reinvigorate the dialogue.

Philip B. Allen is in the Department of Physics, State University of New York at Stony Brook, Stony Brook, New York 11794-3800, USA.
e-mail: philip.allen@sunysb.edu

1. Lanzara, A. *et al.* *Nature* **412**, 510–514 (2001).
2. Wollman, D. A. *et al.* *Phys. Rev. Lett.* **71**, 2134–2137 (1993).
3. Valla, T. *et al.* *Science* **285**, 2110–2113 (1999).
4. Hengsberger, M. *et al.* *Phys. Rev. Lett.* **83**, 592–595 (1999).
5. Lashell, S., Jensen, E. & Balasubramanian, T. *Phys. Rev. B* **61**, 2371–2374 (2000).
6. Valla, T. *et al.* *Phys. Rev. Lett.* **83**, 2085–2088 (1999).
7. Allen, P. B. & Dynes, R. C. *Phys. Rev. B* **12**, 905–922 (1975).
8. Kulic, M. L. & Dolgov, O. V. *Physica C* **341**, 111–112 (2000).
9. Maksimov, E. G. *Usp. Fiz. Nauk* **170**, 1033–1061 (2000).
10. Weger, M. & Peter, M. *Physica C* **318**, 252–258 (1999).
11. Phillips, J. C. *Phil. Mag.* **B 81**, 35–54 (2001).
12. Cappelluti, E. *et al.* *Phys. Rev. Lett.* **85**, 4771–4774 (2000).
13. Paci, P. *et al.* *Int. J. Mod. Phys. B* **14**, 2982–2987 (2000).
14. Emin, D. *Phys. Rev. B* **49**, 9157–9167 (1994).
15. Emin, D. J. *Supercond.* **8**, 523–526 (1995).
16. Alexandrov, A. S. *Physica C* **341**, 107–110 (2000).
17. Rodriguez, C. O. *et al.* *Phys. Rev. B* **42**, 2692–2695 (1990).
18. Cohen, R. E. *et al.* *Phys. Rev. Lett.* **64**, 2575–2578 (1990).
19. Allen, P. B. *Comm. Cond. Mat. Phys.* **15**, 327–353 (1992).
20. Friedmann, T. A. *et al.* *Phys. Rev. B* **42**, 6217–6221 (1990).
21. Allen, P. B., Pickett, W. E. & Krakauer, H. *Phys. Rev. B* **37**, 7482–7490 (1988).
22. Mazin, I. I. & Dolgov, O. V. *Phys. Rev. B* **45**, 2509–2511 (1992).

Daedalus

Emotional education

Last week Daedalus proposed to map the mental world of love. By scanning the brains of lovers of all types, and correlating their brain states with their verbal descriptions of them, he hoped to devise objective definitions of words such as ‘desire’, ‘infatuation’ and ‘being in love’, whose wild and inconsistent usage causes vast confusion between the sexes.

He is now generalizing this project. Nobody can truly know what anyone else is feeling. But brain-scan technology should at least allow us to draw consistent mental maps. For example, some medics now present patients with a long list of words for pain — dull, sharp, nagging, stabbing, burning, and so on — to find those words most often used for the pain of specific disorders. By noting which brain regions are activated by pains of a given description, and in what way, Daedalus hopes to draw an abstract ‘pain map’, scaled in character and intensity, and map on to it the most common verbal descriptions. The outcome should be a public vocabulary of private distress, verifiable by brain scanning, in which patients and doctors could at last speak clearly to each other.

Many other private sensations — taste, smell, colour, artistic appeal — would also gain from objective descriptions. But emotion is the most urgent study. For many of us sadly lack ‘emotional intelligence’. Not only do we fail to understand love, we do not recognize the ranges of hate and hope, fear and rage, depression and elation, suspicion and trust, timidity and confidence. By matching people’s verbal descriptions of their mental landscapes with maps of their brain activity, Daedalus hopes to puzzle out an optimal use of common or invented words to describe the normal range of emotions. Anyone could then understand his succession of mental states as a walk on the human mental ‘map’. He could make far better guesses at the states of others, and could explain his feelings in terms that all concerned would recognize. The confusion and friction of social and family life would be replaced, if not by harmony and agreement, at any rate by understanding.

Even science would benefit. Scientists spend their lives in puzzlement, suspended judgement, creative doubt, slowly growing conviction or disbelief — states of mind that students find very perplexing. Any method of teaching these states to the next generation would help to hand on the scientific project continuously to its new and rising stars.

David Jones

Obituary

Viktor Hamburger (1900–2001)

Viktor Hamburger died on 12 June 2001 in St Louis, Missouri, just a few weeks short of his 101st birthday. During his long scientific career, his inquiring mind, sharp intellect and keen observational powers led to key discoveries of the events involved in the formation of the nervous system. His work laid the foundation for identification of the first neurotrophic factor — a protein that enhances the growth and maturation of nerve cells — which in turn led to an appreciation of programmed cell death as a fundamental mechanism in neural development. In the last part of his scientific career, he provided evidence for the embryonic origins of behaviour.

Hamburger received his early training at the University of Freiburg with the acclaimed experimental embryologist Hans Spemann. In 1932 Hamburger, then an assistant professor at Freiburg, travelled to the United States on a Rockefeller Foundation fellowship to work with the embryologist Frank Lillie at the University of Chicago. He brought the method of manipulating embryonic tissues with fine glass needles — which Spemann developed to transplant tissues into salamander embryos — and applied them to the chick embryo. When the Nazis came to power in 1933, Hamburger was relieved of his position in Freiburg, forcing him to remain in the United States, first at Chicago and then at Washington University in St Louis. Thus began a new era in which the chick embryo reigned supreme as the animal model for experimental embryology.

The chick had many advantages over amphibians, commonly used as animal models at the time, such as the possibility of observing embryonic development through a window cut in the egg shell. Adoption of the chick embryo by experimental embryologists was aided greatly by an article published by Hamburger and Howard Hamilton in 1951, entitled “A series of normal stages in the development of the chick embryo” (*Journal of Morphology*), which allowed the ages of embryos to be gauged by their morphology at different times after fertilization. This allowed data from different laboratories to be compared.

Hamburger's landmark work on nerve development, published in 1934, was carried out in Chicago at Lillie's suggestion to repeat an experiment done by his student Marian Shorey. The technique involved transplanting limb buds (limb primordia) onto chick embryos and studying how nerves grow into the



Founder of the field of developmental neuroscience

developing limb. This established the chick embryo as an animal model for experimental embryology and marked the beginning of a new discipline — developmental neuroscience. Hamburger suggested a three-point model for the relationship between developing nerve centres and their targets (such as limbs). The model proposed that specific molecules are transported retrogradely along the nerve fibres that connect targets with nerve centres, to regulate the number of nerve cells within centres. This model laid the foundation for the discovery of nerve growth factor (NGF).

In 1986, the Nobel Prize in Physiology or Medicine was awarded to Rita Levi-Montalcini and Stanley Cohen for the discovery of the first growth factors, one of which was NGF. Although Hamburger had been included in previous prizes for NGF, such as the Horwitz Prize, he was excluded from the Lasker Prize, which immediately preceded the Nobel. Although it seemed entirely appropriate to award the Nobel Prize to Levi-Montalcini and Cohen, the absence of Hamburger appeared odd to the US neuroscience community. It was widely believed that NGF would not have been discovered without Hamburger's 1934 model, followed by his invitation to Levi-Montalcini to join him in 1948 to investigate the relationships between nerve centres and their targets. Moreover, the initial experiments leading to the discovery of NGF, published in 1949 and 1950, were designed by Hamburger and carried out by Levi-Montalcini, after which Cohen was hired to isolate NGF. Historically, the pivotal role of Hamburger in the NGF story seemed clear.

To his friends, students and colleagues, Hamburger was not just a great scientist, but also a great personality, who never took himself too seriously and was humble to a fault. In truth, it may have been his self-deprecating character that led to his exclusion from the Nobel Prize. By his own account, he withdrew from the project once it had become a chemical problem. He left the rest of the story to Cohen and Levi-Montalcini, whom he knew had the talent and drive to complete it. When he was interviewed in 1987 regarding his feelings about the Nobel Prize, his sense of humour came through as he said: “I feel something like Joseph at the manger, always in the background, and his role in the miracle a little bit dubious!”

After closing his laboratory at Washington University at the age of 85, Hamburger turned to his second passion — the history of science. Although he had published occasional articles on historical figures in experimental embryology, he turned his full attention to Hans Spemann and his laboratory. In 1988 he published his magnum opus, *The Foundations of Experimental Embryology: Hans Spemann and the Organizer*. This book quickly became an invaluable resource for developmental biologists, who now had access to the seminal experiments performed by the Spemann school, previously available only in German. But Hamburger did not merely translate the papers; he also described the working hypotheses and experimental designs, and synthesized the findings. This made the work intellectually accessible, and provided a road map to the early work on embryonic induction, still an unsolved problem in developmental biology. Hamburger also provided candid portraits of Spemann and his laboratory, making this important part of history come alive.

He may not have received the Nobel Prize, but Viktor Hamburger founded developmental neuroscience, introduced the chick embryo to experimental embryology, and had a pivotal role in the discovery of the first neurotrophic factor. He was the living link between early experimental embryology and modern developmental biology.

Jean M. Lauder and Ronald Oppenheim
Jean M. Lauder is in the Department of Cell and Developmental Biology, University of North Carolina School of Medicine, Chapel Hill, North Carolina 27599, USA.

e-mail: unclau@med.unc.edu

Ronald Oppenheim is in the Department of Neurobiology and Anatomy, Wake Forest University School of Medicine, Winston Salem, North Carolina 27157, USA.
e-mail: roppenhm@wfubmc.edu

Fossil molar from a Madagascan marsupial

The discovery of a tiny tooth from the Late Cretaceous period has sizeable implications.

Here I report the discovery of an isolated tooth of a marsupial mammal from the Late Cretaceous period in Madagascar. This specimen represents the only known marsupial, fossil or extant, from Madagascar and is the first to be identified with confidence from deposits of Mesozoic age on the prehistoric southern supercontinent of Gondwana.

University of Antananarivo (UA) specimen 8699 is a lower molar that was discovered as part of a joint project by Stony Brook University and the University of Antananarivo in the Mahajanga Basin of northwest Madagascar¹. The tooth was recovered from locality MAD93-35, in the Anembalemba Member of the Maevarano Formation, which has been assigned a Maastrichtian (latest Cretaceous) age².

UA 8699, which is missing a mesiolingual portion of the trigonid, is estimated to have been 3.5 mm long and 2.2 mm wide when complete (Fig. 1). Not enough of the tooth is preserved to warrant the naming of a new species or the allocation of the specimen to a lower taxon, but it exhibits several salient features that are indicative of a relatively derived, tribosphenidan mammal (those with a pestle-and-mortar, 'reversed triangle' type of molar dentition). In particular, the talonid is broadly basined, the angle of the trigonid is acute ($\sim 48^\circ$), there is no distal metacristid, the protocristid is relatively transverse, and the cristid obliqua meets the distal trigonid wall in a relatively buccal position^{3,4}.

It has recently been proposed⁵ that the tribosphenidan molar pattern evolved on at least two separate occasions. Thus, tribosphenidans are divided into a boreosphenidan clade (extant placental mammals, marsupials and their relatives) of presumed Laurasian (northern) origin, and an australosphenidan clade (extant monotremes and their relatives) of presumed Gondwanan (southern) origin. Under this hypothesis, UA 8699 is from a boreosphenidan; it does not have a shelf-like cingulid extending onto the lingual aspect of the crown, a hallmark specialization that unites the Australosphenida.

The fact that UA 8699 represents a marsupial (non-deltatheroidan metatherian), rather than a primitive metatherian (deltatheroidan) or a eutherian (placental mammals and their extinct relatives), is indicated by a range of features^{3,6-11}. First, it has a prominent distobuccal cingulid extending from the base of the hypoconulid towards the hypoconulid apex. Second, the hypoconulid is lingually situated, well away

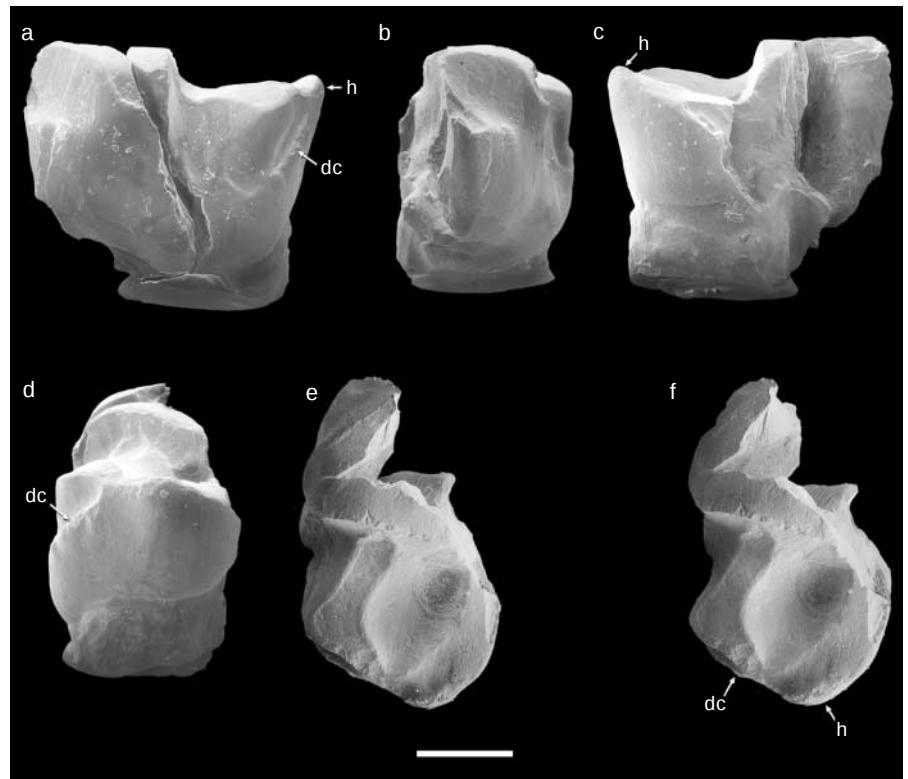


Figure 1 Lower-left molar (UA 8699) of a marsupial mammal from the Late Cretaceous period in Madagascar. Different views of the tooth are shown: **a**, buccal; **b**, mesial; **c**, lingual; **d**, distal; and **e**, **f**, occlusal (stereopair). Abbreviations: h, hypoconulid; dc, distobuccal cingulid. Scale bar, 1 mm.

from the hypoconid (inferred from the configuration of the worn dentine lake). Third, the talonid is at least as broad as the trigonid (as the buccal portion of the trigonid does not extend beyond the buccal portion of the talonid). Fourth, the trigonid is low (although heavily worn and incomplete, the size and shape of the trigonid suggest that the height differential between it and the talonid was not great). Fifth, wear is mainly horizontal, resulting in flat wear platforms with broad bands of exposed dentine.

Gondwanan marsupials have been identified from the Palaeocene to Holocene epochs in South America, the Late Eocene epoch in Antarctica, from the Early Eocene to Oligocene epochs in Africa, and from the Early Eocene to Holocene in Australasia¹². Late Cretaceous marsupials have been reported at several localities in South America but, in all cases, either the age of the strata or the identification of the specimens has been doubtful.

In the context of the generally accepted hypothesis that marsupials originated in the Northern Hemisphere and gained access to Gondwana by dispersal from North to South America¹², the discovery of a marsupial in the latest Cretaceous stage in Madagascar has

implications for hypotheses of biogeography and plate tectonics. It seems that marsupials were broadly distributed on Gondwana by the end of the Cretaceous, having traversed South America and entered Madagascar, presumably through Antarctica. Thus, physical and biotic connections between Madagascar and Antarctica seem to have been maintained until the later stages of the Late Cretaceous, confirming conclusions derived from other elements in the fauna¹.

Ecological displacement or replacement of archaic Gondwanan mammalian fauna (gondwanatheres, pretribosphenic holoherians and australosphenidans) by boreosphenidans may already have been under way by the end of the Cretaceous (as also indicated by the discovery of the placental mammal *Deccanolestes* in India¹³).

Finally, the taxon represented by UA 8699 adds significantly to the accumulating evidence that the ancestors of modern Madagascan mammals, all of which are placental, were not present on the island in the Late Cretaceous¹⁴. As Madagascar was already an island in the Maastrichtian¹⁵, this evidence suggests that the basal stocks of the modern taxa arrived in post-Cretaceous times, and that they must have done so by

crossing a significant marine barrier¹.

David W. Krause

Department of Anatomical Sciences,

Stony Brook University, Stony Brook,

New York 11794-8081, USA

e-mail: dkrause@mail.som.sunysb.edu

- Krause, D. W. et al. *GSA Today* **9**, 1–7 (1999).
- Rogers, R. R., Hartman, J. H. & Krause, D. W. *J. Geol.* **108**, 275–301 (2000).
- Cifelli, R. L. in *Mammal Phylogeny Vol. 1: Mesozoic Differentiation, Multituberculates, Monotremes, Early Therians, and Marsupials* (eds Szalay, F. S., Novacek, M. J. & McKenna, M. C.) 205–215 (Springer, New York, 1993).
- Fox, R. C. *Can. J. Earth Sci.* **12**, 412–442 (1975).
- Luo, Z.-X., Cifelli, R. L. & Kielan-Jaworowska, Z.

- Nature* **409**, 53–57 (2001).
- Cifelli, R. L. & Eaton, J. G. *Nature* **325**, 520–522 (1987).
 - Cifelli, R. L. *Proc. Natl Acad. Sci. USA* **90**, 9413–9416 (1993).
 - Cifelli, R. L. & Muizon, C. *J. Mamm. Evol.* **4**, 241–258 (1997).
 - Clemens, W. A. & Lillegraven, J. A. *Contr. Geol., Univ. Wyoming, Spec. Pap.* **3**, 55–85 (1986).
 - Slaughter, B. H. *Science* **162**, 254–255 (1968).
 - Fox, R. C. & Naylor, B. G. in *Sixth Symposium on Mesozoic Terrestrial Ecosystems and Biota* (eds Sun, A. & Wang, Y.) 247–250 (China Ocean, Beijing, 1995).
 - Woodburne, M. O. & Case, J. A. *J. Mamm. Evol.* **3**, 121–161 (1996).
 - Prasad, G. V. R. & Sahni, A. *Nature* **332**, 638–640 (1994).
 - Masters, J. C., Rayner, R. J. & Tattersall, I. in *Creatures of the Dark: The Nocturnal Prosimians* (eds Alterman, L., Doyle, G. A. & Izard, M. K.) 31–44 (Plenum, New York, 1995).
 - Storey, M. et al. *Science* **267**, 852–855 (1995).

Maternal investment

Sex differences in avian yolk hormone levels

It has been suggested that female birds put more resources into eggs fathered by attractive males by laying larger eggs¹ or by adding more testosterone², but this inference could be undermined if eggs of different sex are provisioned differently, as these studies did not control for sex differences. Here we compare hormone concentrations in the yolks of male and female eggs and find that these are significantly different. Our results indicate that it is premature to conclude that female birds invest more in eggs sired by a preferred male, and raise the possibility that yolk sex steroids may be part of the sex-determining process in birds.

Cunningham and Russell have shown that female mallards (*Anas platyrhynchos*) produce larger eggs when mated to attractive males¹, and Gil *et al.* have found that female zebra finches (*Taeniopygia guttata*) put more androgens into the yolk of eggs laid for attractive males². However, neither study controlled for the possible effect of the sex of the eggs in the treatment groups. Females produce more sons when mated to attractive males³ — for example, it has been shown in zebra finches, using the same colour-band manipulations as Gil *et al.*², that females produce more sons when mated to attractive, red-banded males⁴. If male offspring generally come from bigger eggs or if there are sex differences in the steroid content of eggs, then the results of both studies^{1,2} could be attributable to an effect of sex-ratio differences.

We compared the volume (calculated from Hoyt's formula⁵: volume = $0.51 \times \text{length} \times \text{breadth}^2$) of a large sample of male and female peafowl (*Pavo cristatus*) eggs and found a non-significant tendency for male eggs to be larger (ANCOVA with laying date as a covariate: main effect sex, $F_{1,744} = 3.01$, $P = 0.086$). We also analysed and compared hormone levels in the yolk of female and male peafowl eggs, in which the embryos had been sexed at an early stage by

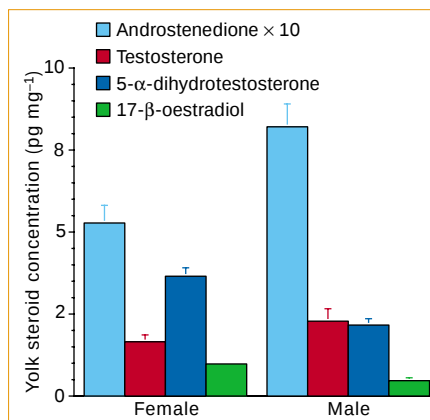


Figure 1 Steroid concentrations (in pg per mg of yolk) in male ($n = 43$) and female ($n = 45$) eggs. Eggs were collected from 46 pens, in which 1 male peacock had been mated to 4 peahens per pen. Eggs were placed in an incubator on day 1 and dissected on day 10 of the 28-day incubation period. A small quantity of embryonic material was collected for sexing and at the same time a sample of yolk was taken for steroid analysis. Yolks were analysed as described¹³ without knowing the sex of the egg. Embryos were sexed by using polyacrylamide-gel electrophoresis of DNA amplified by the polymerase chain reaction. Where possible, one egg of each sex from each pen was analysed for hormonal concentration. We used ANCOVA to compare steroid concentrations between sexes with laying date as a covariate (androstenedione, $F_{1,87} = 10.54$, $P = 0.002$; testosterone, $F_{1,87} = 6.73$, $P = 0.011$; 5-α-dihydrotestosterone, $F_{1,87} = 20.71$, $P < 0.0001$; 17-β-oestradiol, $F_{1,87} = 30.86$, $P < 0.0001$).

using molecular techniques. We found significant differences in overall levels of yolk steroids, including androgens, for eggs of different sex (Fig. 1).

The main source of steroids in egg yolk is the maternal ovary cells that surround the developing egg until it is ovulated⁶. Embryos are not known to secrete sex steroids into the yolk at an early stage, although there is some evidence for the presence of corticosterone of embryonic origin in the egg yolks of tree lizards, *Urosaurus ornatus*, after day 25 of the 36-day incubation period⁷. Also, embryos of each sex may use maternal steroids in a different way, although this has not been investigated in birds. Avian oocytes undergo the first meiotic division in the ovary after completion of yolk deposition, a few hours

before ovulation. This division consigns one of the sex chromosomes to the polar body, effectively determining the sex of the egg. If the ovulated egg is fertilized, it undergoes the second meiotic division and develops into a male or female embryo⁸.

The existence of sex differences in yolk hormone levels suggests that female birds provide eggs destined to carry male and female embryos with different amounts of steroids. It is unlikely that females can tell the sex of the egg when providing the yolk with hormones, as sex is determined after the yolk is formed. Instead, we propose that maternal steroids influence sex-chromosome segregation at the first meiotic division and are thus part of the sex-determining mechanism^{9–11}. Maternally derived yolk hormones are associated with offspring sex ratio in painted turtles (*Chrysemys picta*)¹², in which sex is environmentally determined.

If zebra-finch eggs show similar sex-related differences in the amount of hormones they contain, then the increased testosterone levels reported by Gil *et al.*², in groups in which females were mated to attractive males, may be a result of the presence of more male eggs in these clutches⁴. We therefore question whether there is differential investment in hormones by females in the eggs of attractive males, as the results may be explained by females biasing the sex ratio of their offspring³.

Marion Petrie*, Hubert Schwabl†, Nanna Brande-Lavridsen‡, Terry Burke‡

*Evolution and Behaviour Research Group, Department of Psychology, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK
e-mail: marion.petrie@ncl.ac.uk

†Center of Reproductive Biology, School of Biological Sciences, Washington State University, Pullman, Washington 99164, USA

‡Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

- Cunningham, E. J. A. & Russell, A. F. *Nature* **404**, 74–75 (2000).
- Gil, D., Graves, J., Hazon, N. & Wells, A. *Science* **286**, 126–128 (1999).
- Sheldon, B. C., Andersson, S., Griffith, S. C., Ornborg, J. & Sendecka, J. *Nature* **402**, 874–877 (1999).
- Burley, N. *Evolution* **40**, 1191–1206 (1986).
- Hoyt, D. F. *Auk* **96**, 73–77 (1979).
- Schwabl, H. in *Perspectives in Avian Endocrinology* (eds Harvey, S. & Etches, R. J.) 3–13 (J. Endocrinol., Bristol, 1997).
- Jennings, D. H., Weiss, S. L. & Moore, M. C. *Am. Zool.* (in the press).
- Sturkie, P. D. *Avian Physiology* (Springer, New York, 1986).
- Bogart, M. H. *J. Theor. Biol.* **128**, 349–357 (1987).
- Crews, D., Cantu, A. R., Rhen, T. & Vohra, R. *Gen. Comp. Endocrinol.* **102**, 317–326 (1996).
- Krackow, S. *Johannesburg: Bird Life South Africa. Proc. 22nd Int. Ornithol. Congr.* (Durban, 1999).
- Bowden, R. M., Ewert, M. A. & Nelson, C. E. *Proc. R. Soc. Lond. B* **267**, 1745–1749 (2000).
- Schwabl, H. *Proc. Natl Acad. Sci. USA* **90**, 11446–11450 (1993).

Cunningham and Russell reply — Female birds may invest more in breeding attempts when paired with preferred males and may consequently produce chicks that are in better condition¹. But if females favour one sex when mated to preferred males, either

Table 1 Factors affecting egg size and sex

Response term	Explanatory term	Statistic	d.f.	P value
Embryo sex	Male rank	1.00	1	0.32
Egg volume	Embryo sex	4.27	1	0.04
	Male rank	6.69	1	0.01
	Male rank $\times$ offspring sex	0.4	1	0.53
Chick weight	Male rank	6.29	1	0.01
	Embryo sex	0.07	1	0.78
	Male rank $\times$ offspring sex	<0.01	1	0.97
Per cent males (2nd clutch)	Per cent males (1st clutch)	2.78	1,15	0.03
Egg volume (2nd clutch)	Egg volume (1st clutch)	3.23	1,15	0.006
Per cent males	Mean egg volume	0.03	1	0.86

Effects of embryo sex and male rank on egg volume and chick weight were investigated using residual maximum likelihood models. There were no differences in chick size at hatching ($\chi^2 = 0.03$, d.f. = 1, $P = 0.47$). Generalized linear mixed models with binomial error structure and logit link function were used to investigate the effects of male rank and a female's average egg volume on the proportion of male eggs laid. Repeated measures within females were controlled for in these analyses ($n = 2$ clutches, 16 females). General linear models were used to investigate the repeatability in sex ratio and egg volume (Genstat 5.4.1, Lawes Agricultural Trust, IACR Rothamsted; 1998) (E.J.A.C., A.F.R., K. Orr, R. Griffiths and D. J. Ross, unpublished results).

in quantity^{2,3} or quality, could this explain these differences in investment? We have shown that female mallards lay larger eggs for preferred males but do not produce more sons¹. This increased investment is not directed at one particular sex, and here we point out the importance of distinguishing between differential investment in the sexes *per se*, as suggested by Petrie *et al.*, and differential investment in the sexes for different males.

Table 1 (top) shows that, within any breeding attempt, female mallards lay larger eggs for male embryos. However, this does not explain the increased investment for preferred males between different breeding attempts. Also, larger eggs for preferred males then produce heavier chicks, irrespective of their sex. Hence, in mallards, both sexes benefit from their mother's increased investment with preferred males.

Maternal characteristics seem to have a strong effect on the sex of their offspring (Table 1, bottom). Females that produce a high proportion of males in their first clutch also produce a high proportion of males in their next clutch, regardless of their partner's rank. Females are also consistent in their egg size after controlling for differential investment for different males, but females that generally produce larger eggs do not produce more males.

Whether or not differences in hormonal levels found in clutches sired by different males translate into 'differential investment' is more complex. Overall increases in hormonal level would suggest differential investment. Alternatively, differential allocation of hormones within a clutch could simply be a consequence of biasing the sex ratio in favour of a particular sex. But if the favoured sex is more costly to produce, then this would still represent a form of differential investment.

The key issue is the effect of this bias in investment. Does biasing the sex ratio of offspring increase the overall viability of the clutch because the preferred sex is more

likely to survive? Or does the less preferred sex suffer in the trade-off to the detriment of their survival? Females may bias investment in the sexes for many different reasons, so understanding how they influence the success of their sons and daughters will also be essential in explaining why birds sired by different males differ in their success.

Finally, could maternal provisioning of egg hormone levels be linked to the sex-determining process, as suggested by Petrie *et al.*? First we need to find out whether male phenotype can influence hormonal levels in the maternal body and ovary, and then whether manipulating hormonal levels within the female body and ovary can bias production of the sexes.

Emma J. A. Cunningham, Andrew F. Russell

Department of Zoology, University of Cambridge,
Downing Street, Cambridge CB2 3EJ, UK

e-mail: ejac3@hermes.cam.ac.uk

1. Cunningham, E. J. A. & Russell, A. F. *Nature* **404**, 74–77 (2000).
2. Burley, N. *Evolution* **40**, 1191–1206 (1986).
3. Sheldon, B. C., Andersson, S., Griffiths, S. C., Ornborg, J. & Sendecka, J. *Nature* **402**, 874–877 (1999).

Ecology

Global amphibian population declines

The decline and disappearance of relatively undisturbed populations of amphibians in several high-altitude regions since the 1970s suggests that they may have suffered a global decline, perhaps with a common cause or causes^{1–3}. Houlahan *et al.*⁴ examined means of trends for 936 amphibian populations and concluded that global declines began in the late 1950s, peaked in the 1960s, and have continued at a reduced rate since. Here we re-analyse their data using a method that accounts for the sampling of different populations over different time periods, and find evidence of a mean global decline in monitored populations only in the 1990s. However it is calculated, the global mean not only masks

substantial spatial and temporal variation in population trends and sampling effort, but also fails to distinguish between a global decline with global causes and the cumulative effects of local declines with local causes.

The first analytical method used by Houlahan *et al.* evaluated patterns in ΔN , the annual change in abundance within populations, where

$$\Delta N_t = \log(N+1)_{t+1} - \log(N+1)_t$$

They calculated annual mean changes by averaging ΔN_t over populations with recorded abundances in years t and $t+1$. The temporal pattern of this mean is misleading because each year includes different populations. For example, consider one population studied in years 1, 2 and 3 with $\Delta N_1 = 0.1$, $\Delta N_2 = 0.3$, and another studied in years 2, 3 and 4 with $\Delta N_2 = -0.5$, $\Delta N_3 = -0.3$. Both populations are doing better with time, but the arithmetical averages for the three years, $\bar{\Delta N}_1 = 0.1$, $\bar{\Delta N}_2 = -0.1$, $\bar{\Delta N}_3 = -0.3$, indicate the opposite. The correct approach to estimating $\bar{\Delta N}_t$ uses least-squares means⁵, which estimate the yearly mean averaged over all populations. The least-squares means of ΔN_t , -0.3 , -0.1 and 0.1 , correctly represent the observed trends.

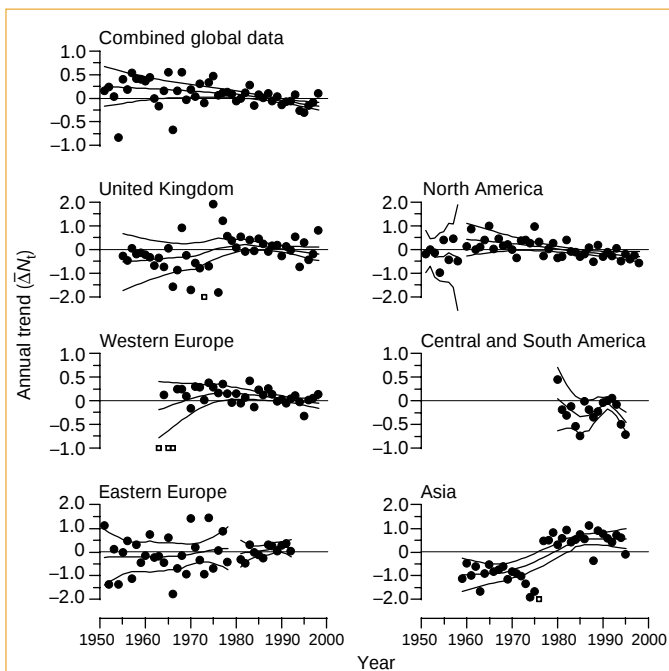
We used least-squares means to estimate annual mean trends for the combined global data and for several geographical regions⁴ (Fig. 1). The global mean trend was significantly positive during 1964–81, indicating that, on average, monitored amphibian populations increased over this period. From 1990 onwards, the trend was significantly and increasingly negative, suggesting a global decline. This decline began more than a decade later than is generally accepted^{1,2}, a quarter of a century later than Houlahan *et al.*⁴ found, and at around the time that concerns about it were first expressed³.

The global mean does not reflect a single worldwide trend. For regions where data allowed separate analysis, trends of the 1990s were significantly negative only in North America and in Central and South America (Fig. 1), where concerns have been raised over amphibian declines^{1–3,6,7}. The trend for Asia was significantly negative during 1959–75 and significantly positive after 1983.

The second analysis carried out by Houlahan *et al.* identified significantly more negative than positive correlations of population size with time, and they interpret this as evidence of a global decline. However, this pattern is expected for many amphibians in which recruitment is more variable than survival, and the exact expectation depends on the population biology of each species¹. It is therefore impossible to establish a correct null hypothesis for the global database.

Extrapolating these results beyond the particular populations studied is tenuous.

Figure 1 Yearly least-squares means for ΔN_t , estimated using the statistical-analysis-system procedure mixed⁵ for the combined global data and for different regions. Only those populations with data for three or more consecutive years are included, to allow adjustment for first-order autocorrelation. Data points that fall outside the axes are indicated by hollow squares; the values for these are: UK 1973, -2.12; western Europe: 1963, -2.37; 1965, -1.28; 1966, -2.63; Asia 1976, -3.39. Trend lines are non-parametric regressions using LOESS⁸, weighting points by $1/s.e.^2$; approximate 95% confidence limits for LOESS regressions are indicated by outer lines.



Sampling intensity varied widely within and among regions and over time⁴, leading to variation in the power and generality of analyses. Amphibian population declines in less modified habitats are associated with high altitudes^{1,2}, but elevation was not considered in the analyses⁴. Furthermore, the populations included in the data set were studied for many reasons, including concern over possible decline or human impact, and were not a representative random sample of the global amphibian fauna. Determining the true nature and extent of global trends in amphibian populations will require the collection of more data and more detailed analysis.

Ross A. Alford*, **Philip M. Dixon†**,
Joseph H. K. Pechmann‡

*School of Tropical Biology, James Cook University,
Townsville, Queensland 4811, Australia
e-mail: ross.alford@jcu.edu.au

†Department of Statistics, Iowa State University,
125 Snedecor Hall, Ames, Iowa 50011-1210, USA

‡Department of Biological Sciences, University of
New Orleans, New Orleans, Louisiana 70148, USA

1. Alford, R. A. & Richards, S. J. *Annu. Rev. Ecol. Syst.* **30**, 133–165 (1999).
2. Waldman, B. & Tocher, M. in *Behavioural Ecology and Conservation Biology* (ed. Caro, T.) 394–443 (Oxford Univ. Press, New York, 1998).
3. Wake, D. B. & Morowitz, H. J. *Alytes* **9**, 33–42 (1991).
4. Houlahan, J. E., Findlay, C. S., Schmidt, B. R., Meyer, A. H. & Kuzmin, S. L. *Nature* **404**, 752–755 (2000).
5. Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. *SAS System for Mixed Models* (Cary, North Carolina, 1996).
6. Berger, L. et al. *Proc. Natl Acad. Sci. USA* **95**, 9031–9036 (1998).
7. Pounds, J. A., Fogden, M. P. L., Savage, J. M. & Gorman, G. C. *Conserv. Biol.* **11**, 1307–1322 (1997).
8. Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A. *Graphical Methods for Data Analysis* (Wadsworth, Belmont, California, 1983).

Houlahan et al. reply — Alford et al. address several questions related to the biological and statistical analysis of declines in global

amphibian populations. They argue that, by emphasizing the global mean, we have masked spatial and temporal variation in amphibian population trends. Admittedly, information is lost when any summary statistic is used, but global amphibian declines should not be inferred by estimating missing values. Furthermore, they contend that we do not distinguish between a global decline with global causes and the cumulative effects of local declines with local causes. But we did not address the issue of causation: we reported widespread declines in extant, mostly lowland populations, whereas the recent focus has been on extinctions at high-altitude sites^{1,2}.

Alford et al. disagree with our statistical analysis, and re-analyse our data set using a method that involves estimating about 38,000 unobserved values, some of which are not biologically feasible. The data used in both analyses are annual population growth rates ($dN_{t,p}$, where t is the year and p is the population) from about 900 amphibian-population time series during 1950–97. None of the populations spans the entire period, so the data are a roughly 900×47 matrix of $dN_{t,p}$ with about 5,000 observed values and about 38,000 empty cells.

The least-squares means approach used by Alford et al. tacitly estimates the missing values ($dN_{t=1,p=2}$ and $dN_{t=3,p=1}$) by assuming that the difference between $dN_{t=1,p=1}$ and $dN_{t=2,p=1}$ can be used to infer $dN_{t=1,p=2}$ (and, in the same manner, $dN_{t=3,p=1}$); $dN_{t=2,p=1} - dN_{t=1,p=1} = 0.3 - 0.1 = 0.2$. Thus, subtract 0.2 from $dN_{t=2,p=2}$ to get $dN_{t=1,p=2} = -0.5 - 0.2 = -0.7$. The mean of $dN_{t=1,p=1}$ (+0.1) and $dN_{t=1,p=2}$ (-0.7) is -0.3, the least-squares means value calculated by Alford et al. for year 1.

Estimating 38,000 missing values on the

basis of 5,000 observed values is problematic, and is made worse by ignoring the fact that population sizes are associated with dN values. According to the least-squares means approach, populations can grow without limit and negative dN values can be estimated for extinct populations. Consider again two populations: the sizes of population 1 are $N_1=99$, $N_2=31$ and $N_3=0$ with $dN_{t=1,p=1}=-0.5$ and $dN_{t=2,p=1}=-1.5$. For population 2, $N_2=0$, $N_3=49$ and $N_4=99$, and $dN_{t=2,p=2}=1.7$ and $dN_{t=3,p=2}=0.3$. The estimation procedure used by Alford et al. yields $\Delta N_{t=3,p=1}=-2.9$, $\Delta N_{t=1,p=2}=2.7$, implying that population sizes were negative for population 1 in year 4 and for population 2 in year 1. By randomly selecting 20 populations and estimating the missing values for the 1950–97 period, we found that about 20% of the estimated dN values were associated with negative population size.

Alford et al. also suggest that the ratio of positive correlations of population size with time to negative correlations is not meaningful. They claim^{3,4} that stationary amphibian population dynamics are characterized by many small, negative dN values and few large, positive dN values. However, in this data set, the average negative change (-0.299 ± 0.007 s.e. ($N=2,464$)) is virtually identical to the average positive change (0.302 ± 0.007 s.e. ($N=2,270$)).

Whether these 936 amphibian populations are representative of the global amphibian fauna of all extant (or recently extinct) populations is unknown and cannot be determined, so the magnitude and direction of any bias cannot be estimated. The choice is therefore whether or not to use the data available to address an important conservation issue. We opted to do so and conclude that amphibians have been and are still declining.

Jeff E. Houlahan*, **C. Scott Findlay†**,
Andrea H. Meyer‡, **Sergius L. Kuzmin§**,
Benedikt R. Schmidt¶

*Department of Biology, University of Ottawa,
30 Marie Curie, Ottawa, Ontario K1N 6N5, Canada
e-mail: jeffhoul@science.uottawa.ca

†Center for Research on Environment–Economy
Interactions, University of Ottawa, 5 Calixa
Lavalée, Ottawa, Ontario K1N 6N5, Canada

‡Swiss Federal Statistical Office, Sektion
Hochschulen und Wissenschaft, Espace de
l'Europe 10, CH-2010 Neuchâtel, Switzerland

§Institute of Ecology and Evolution, Russian
Academy of Sciences, Moscow 117071, Russia

¶Zoologisches Institut, University of Zürich,
Winterthurerstrasse 190, CH-8057 Zürich,
Switzerland

1. Laurance, W. F., McDonald, K. R. & Speare, R. *Conserv. Biol.* **10**, 406–413 (1996).
2. Lips, K. R. *Conserv. Biol.* **13**, 117–125 (1999).
3. Alford, R. A. & Richards, S. J. *Annu. Rev. Ecol. Syst.* **30**, 133–165 (1999).
4. Pechmann, J. H. K. & Wilbur, H. M. *Herpetologica* **50**, 65–84 (1994).

The Earth's mantle

George R. Helffrich*† & Bernard J. Wood†

* Earth and Planetary Science, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-Ku, Tokyo 152-8551, Japan

† Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, UK

Seismological images of the Earth's mantle reveal three distinct changes in velocity structure, at depths of 410, 660 and 2,700 km. The first two are best explained by mineral phase transformations, whereas the third—the D" layer—probably reflects a change in chemical composition and thermal structure. Tomographic images of cold slabs in the lower mantle, the displacements of the 410-km and 660-km discontinuities around subduction zones, and the occurrence of small-scale heterogeneities in the lower mantle all indicate that subducted material penetrates the deep mantle, implying whole-mantle convection. In contrast, geochemical analyses of the basaltic products of mantle melting are frequently used to infer that mantle convection is layered, with the deeper mantle largely isolated from the upper mantle. We show that geochemical, seismological and heat-flow data are all consistent with whole-mantle convection provided that the observed heterogeneities are remnants of recycled oceanic and continental crust that make up about 16 and 0.3 per cent, respectively, of mantle volume.

The Earth's mantle comprises 82% of its volume and 65% of its mass. It constitutes virtually all of the silicate part of the Earth, extending from the base of the crust (0.6% of Earth's silicate mass) to the top of the metallic core (Fig. 1). When the core segregated from the silicate and gas of the proto-Earth, it incorporated high concentrations of the **siderophile** elements, leaving **lithophile** elements in the silicate mantle. (Words in bold are explained in the glossary; see Box 1.) Thus, the current composition of the mantle has core formation imprinted on it—as pronounced depletions in, for example, Fe, Ni, S, W, Pt, Au and Pb relative to the **chondritic meteorites**^{1,2}, which are used to constrain the composition of the whole Earth. Owing to the fractionation of lithophile radioactive parent isotopes such as ²³⁸U and ¹⁸²Hf from their siderophile daughters ²⁰⁶Pb and ¹⁸²W, core formation can be dated as the time at which the evolution of the isotopic compositions of Pb and W diverged from the meteorite trend. The result (about 4.5 Gyr ago) corresponds to 50–100 Myr

after the formation of the oldest meteorite bodies in the Solar System^{3,4}.

Mantle rocks that occur occasionally at the surface, either as tectonic fragments (kilometre scale) or as inclusions in explosive eruptives (centimetre scale) are predominantly **peridotites**. One of the main questions facing the Earth sciences is whether a peridotite composition, representative of the upper 150 km of the mantle, can also be assumed to represent the remainder of the mantle down to 2,900 km depth. Many geochemical data suggest not. For example, the current heat flux at the Earth's surface is about 44 TW (44×10^{12} W), most of which can be reasonably attributed to radioactive decay of K, U and Th in the mantle⁵. The upper-mantle source region of mid-ocean ridge **basalt** (MORB) is depleted in these elements, however, and only produces 2–6 TW. The simplest explanation of the shortfall is that there is a lower layer enriched in the heat-producing elements that is only sporadically involved in the production of surface rocks^{6–8}. Similarly, the flux of the

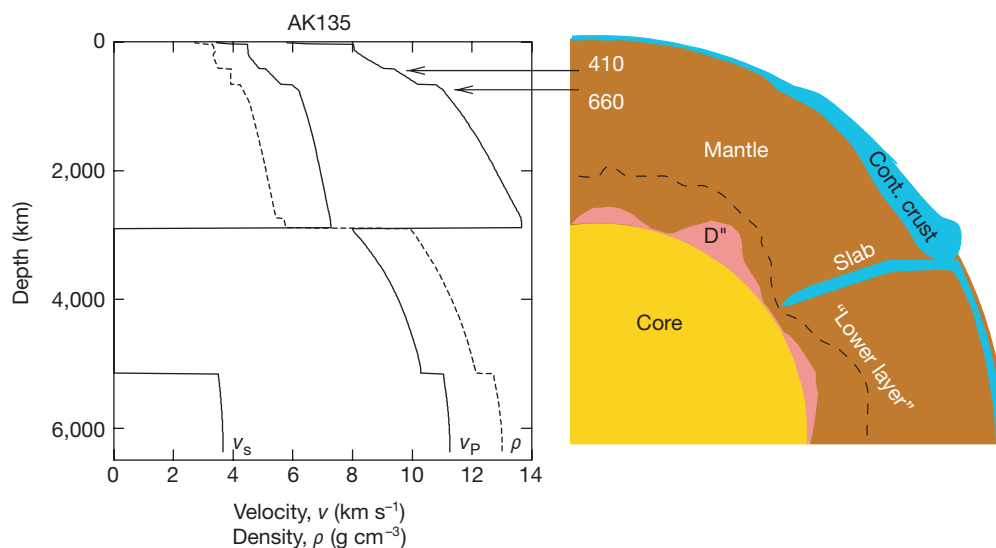


Figure 1 Speeds of seismic waves in the Earth, showing the major discontinuities at 410 and 660 km depth, and the D" layer at the base of the mantle. Also shown (right) is a cut-away view of the Earth, showing a region of continental crust and a subducted lithospheric slab extending into the lower mantle. The D" layer is irregular in thickness. A hypothetical "lower layer"⁵ is shown with a dashed line.

radioactive products ^{40}Ar (from K) and ^4He (from U, Th) into the atmosphere should be predictable from the surface heat-flow and the approximate abundances of K, U and Th in the bulk Earth. In fact, about 50% of the ^{40}Ar produced over the age of the Earth is missing from the atmosphere⁹, while the flux of ^4He from the oceans is only 5% of that predicted from oceanic heat flow¹⁰. These observations again lead naturally to the concept of a region in the deep Earth, rich in the products of radioactive decay, which exchanges heat but little mass with the convecting upper mantle (Fig. 1).

The object of this Review is initially to summarize the data relating to the structure and composition of the mantle with an emphasis on those observations in favour of, or against, chemical layering. We will then present some possible ways in which conflicting evidence might be reconciled, and suggest some future directions that will help resolve the remaining questions.

Mantle structure

The principal data relating to the deep interior are those of seismology. Global one-dimensional seismic models¹¹ (Fig. 1) clearly show the main features that constrain the mineralogical and chemical constitution of the deep interior. More detailed regional studies provide important additional information. Figure 1 shows that there are two main seismic discontinuities in the mantle, at 410 and 660 km depth. In the lower mantle, below 660 km, seismic velocities increase monotonically with depth until the D'' region is reached at about 200 km above the core–mantle boundary. The latter is a region of low gradient in seismic wave speeds, and contains large-scale low-velocity structures and regional discontinuities¹². Although the mantle discontinuities (410, 660 and D'') are all, potentially, due to chemical layering of the mantle, pressure-induced phase transformations in peridotite are more plausible explanations for the two shallower discontinuities.

Phase transformations. Bernal¹³ was the first to propose that rapid increases in seismic velocity in the mantle might be due to phase transformations rather than a change in composition. Experiments in the mid-1960s¹⁴ showed that the olivine component of peridotite undergoes successive pressure-dependent transformations to the

spinel structure (ringwoodite), and ultimately breaks down to form (Mg,Fe)SiO₃ perovskite plus (Mg,Fe)O (refs 15,16):

$(\text{Mg, Fe})_2\text{SiO}_4 = (\text{Mg, Fe})_2\text{SiO}_4$		Pressure, 13–14 GPa; depth, 410 km
Olivine	Wadsleyite	
$(\text{Mg, Fe})_2\text{SiO}_4 = (\text{Mg, Fe})_2\text{SiO}_4$		Pressure, 18 GPa; depth, 520 km
Wadsleyite	Ringwoodite	
$(\text{Mg, Fe})_2\text{SiO}_4 = (\text{Mg, Fe})\text{SiO}_3 + (\text{Mg, Fe})\text{O}$		Pressure, 23 GPa, depth, 660 km
Ringwoodite	Perovskite	Magnesiowüstite

The transformations of olivine to wadsleyite, and ringwoodite to perovskite (+ oxide), clearly correlate with the two major global seismic discontinuities, and must generate at least some part of the seismic 'signal'. But do phase transformations at constant peridotite bulk composition completely explain the seismic phenomena, or are chemical discontinuities also required? Consider the responses of the three transformations to changes in temperature. Figure 2 (ref. 17) shows a pressure–temperature (*P*–*T*) diagram for pure Mg₂SiO₄ that illustrates the main points. The transformation of olivine to wadsleyite has a positive slope, with *dP/dT* of approximately +3 MPa K^{−1}. In contrast, the breakdown of ringwoodite to perovskite plus magnesiowüstite has a negative slope of −2 MPa K^{−1}. Thus, if the 410- and 660-km discontinuities are entirely due to these phase transformations, regions of abnormally low temperature such as subduction zones should correspond to elevation of the '410' to lower depths and depression of the '660' to greater depths. We address the testing of this prediction in the next section.

A more direct approach to correlating mineralogy and seismic properties is through the elastic properties of the whole mineral assemblage. This requires consideration of transformations in the non-olivine component (30%) through the transition zone (410–660 km) and into the lower mantle.

Phase changes in the pyroxene and garnet components of mantle peridotite are gradual, and lead to changes in slope of the curves of seismic velocity versus depth rather than to discrete discontinuities. The principal transformations involve, first, dissolution of pyroxene into the garnet structure at pressures corresponding to 350–500 km depth^{18,19}. Then, at about 580 km depth, a few per cent of CaSiO₃ perovskite begins to exsolve from the garnet, which at that point constitutes about 30% of a biminerally (garnet + ringwoodite) mantle:

Box 1

Glossary

Basalt A rock made by 5–20% melting of peridotite. It is produced under mid-ocean ridges (mid-ocean ridge basalt, MORB) by upwelling of hot mantle. Ocean island basalts (OIBs) are products of similar processes under ocean islands.

Chondritic meteorites A class of meteorites that record the early history of the Solar System. They are thought to represent primitive planetary materials.

EM-1, EM-2 basalts OIBs having isotopic similarities to recycled continental materials.

HIMU basalts OIBs having isotopic similarities to recycled oceanic crust.

Incompatible/compatible When rock starts to melt, some elements (incompatible) are concentrated in the liquid silicate while others (compatible) remain in the solid minerals.

Lithophile/siderophile Elements that dissolve readily in metallic iron under moderately reducing conditions are siderophile, and are concentrated in the core. Those that more readily bond with oxygen are lithophile, and are concentrated in the silicate Earth.

Peridotites Rocks made up predominantly of olivine (Mg,Fe)₂SiO₄, with lesser amounts of orthopyroxene (Mg,Fe)SiO₃ and clinopyroxene Ca(Mg,Fe)Si₂O₆. Although rarely found at the surface, they appear to be the dominant rock type of the mantle.

Refractory Refractory elements condense from gas to solid at high temperatures.

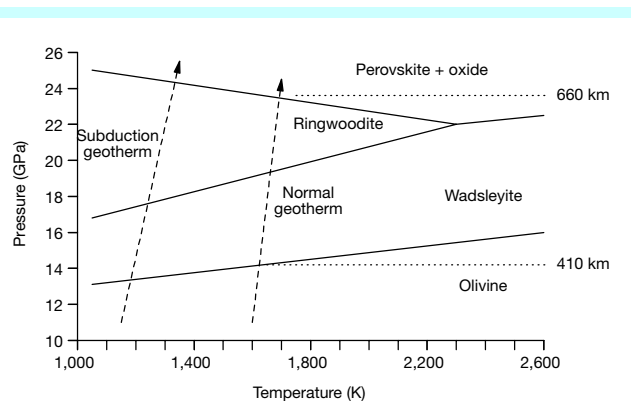
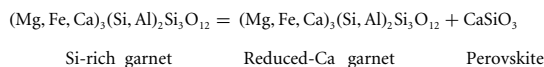
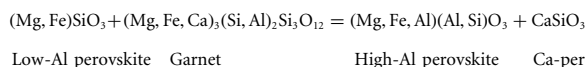


Figure 2 Schematic phase diagram of Mg₂SiO₄ olivine in mantle peridotite. Solid lines show the phase stability regions for olivine, wadsleyite, ringwoodite, and perovskite + oxide. Dashed lines show approximate temperature increase with depth in the mantle and inside a subduction zone. Cooler temperatures in subduction zones move the phase transitions either to shallower depths (olivine → wadsleyite) or deeper depths (ringwoodite → perovskite + oxide).



Finally, between 23 and 26 GPa (660 and 750 km depth), garnet dissolves into the perovskite phase that had initially been produced by breakdown of ringwoodite²⁰:



Predicted mineralogy should enable calculation of seismic wave speeds, which, if the chemical composition is correct, should match the average one-dimensional wave speed structure. Unfortunately, however, individual elastic constants of the high-pressure minerals have not yet been measured at pressures (13–23 GPa) and temperatures (1,400–1,800 °C) relevant to the Earth's interior. Values of bulk modulus K and density ρ are available, even under lower-mantle conditions^{21,22}, so $\Phi = K_S/\rho = v_P^2 - (4/3)v_S^2$, rather than the individual seismic velocities v_P and v_S , provides a well-constrained comparison between seismology and mineralogical properties. (Here v_P and v_S are the velocities of P and S waves, respectively.) Results of such calculations are disappointing, however. The data are not accurate enough to discriminate between a deep mantle of peridotitic composition and one which is enriched in Fe or Si, as has been proposed by numerous authors^{23–25}.

Seismological constraints. Seismological results stimulated most of the ideas on mantle structure discussed above. Historically, radial variations in seismic velocities were derived from analysis of long-period seismic-wave travel times^{26–28}. More recently, seismologists have used both short- and long-period data for tomographic models of wave-speed variations^{29–31}. Additionally, regional short-period seismic arrays comprising hundreds of instruments have revolutionized the study of small-scale mantle structures—such as the widths of the 410- and 660-km discontinuities³². These studies complement the long-period global results by providing detail at small, 10-km, length scales.

The mineralogical reactions responsible for the 410- and 660-km discontinuities have characteristic pressure–temperature slopes ($dP/dT = \Delta S/\Delta V$ where ΔS and ΔV refer to entropy and volume changes of reaction, respectively.) (Fig. 2) that can, in conjunction with short-period seismic studies, be used to constrain mantle properties^{33,34}. In the vicinity of a predictable change in mantle temperature structure a phase transformation must change depth, leading to predictable and testable discontinuity topography. For example, the temperature difference (~ 700 K) between the interior of a subducted lithospheric slab and the surrounding mantle³⁵ should lead to displacements of approximately -60 km and $+30$ km of the 410- and 660-km discontinuities, respectively, if they are solely due to the isochemical transformations of olivine to wadsleyite and ringwoodite to perovskite + magnesiowüstite³⁶. A study of the depth of the 410-km discontinuity³⁷ in the Izu-Bonin subduction zone, based on data from short-period arrays, indicated maximum uplift to about 350 km. Detailed study of the 660-km discontinuity in the same region revealed regional depression of about $+30$ km, and no further mantle features below this depressed transition for a further 300 km (ref. 38). Near subduction zones, the width of the depressed '660 km' transition is a few hundred kilometres^{37–39}, approximating the dimensions of the 'thermal halo' around a cool slab. The depth and width of the depression is thus in agreement with the simple phase-transformation model of the 660-km discontinuity, and contrary to the predicted dynamical behaviour of subducted material intruding into (or deflected by) a compositionally different layer^{40,41}. In the latter case, depression of the discontinuity is predicted to be of the order of several hundred kilometres. The extent of discontinuity topography near subduction zones is thus consistent with isochemical phase changes, and supports the hypothesis that there are no compositional changes at 410 or 660 km depth. The results are also consistent with

penetration of slabs into the lower mantle and with deep recycling of former lithospheric material.

Global seismic tomographic images, based on both short- and long-period data, also strongly suggest slab penetration into the lower mantle. A variety of modern studies show narrow features of high seismic velocity extending from sites of present-day subduction deep into the lower mantle^{29,30,42}. The implication of these seismological observations, in combination with dynamical modelling of mantle mixing^{43,44}, is that there is no major change of chemical composition between the upper and lower mantles. But in contrast to the well-resolved high-velocity structures extending into the lower mantle, the diffuse form of warm low-velocity structures in tomographic images do not as yet convincingly delineate return flow from the lower to the upper mantle.

An important feature of the lower mantle that long-period tomographic studies do not resolve is small-scale elastic heterogeneity, which has been discovered using scattered waves^{45–48}. These studies observed bodies distributed throughout the lower mantle that were smaller than 10 km in size, and whose wave speeds varied by 1% or more from that of the ambient mantle. In the few cases where individual scatterers were identifiable, they had seismic velocities at least 4% slower than their surroundings⁴⁸ and arrayed themselves in linear structures resembling heterogeneities stretched by convective stirring in the mantle^{49–51}. These 'scatterers' must be chemical rather than thermal heterogeneities because they are too small (assuming thermal diffusivity of $1 \times 10^{-6} \text{ m}^2 \text{ s}^{-1}$; ref. 52) to maintain a temperature difference from the mantle for longer than about 200,000 years. Given the tomographic evidence for deep subduction, the most likely explanation is that they are remnants of lithospheric slabs.

Substantial changes in seismic wave speeds occur locally in the lowermost part of the mantle, D'' . This approximately 200-km-thick region varies in thickness, and contains discrete discontinuities⁵³, scattering features^{46,54–56}, and possibly, thin low-velocity layers against the core–mantle boundary^{57,58}. D'' is also anisotropic, and the variations in P-wave and S-wave speeds are uncorrelated^{59,60}. In combination, these features suggest that D'' is a region with pronounced lateral chemical heterogeneity that differs chemically from the overlying mantle.

These seismological properties of mantle structure may be summarized in a map (Fig. 3) of the angular correlation function of seismic heterogeneity provided by analysing recent mantle tomographic studies⁶¹. The angular correlation function reveals the lateral extent of correlated variations in wave speed at a given depth in the mantle. Lateral heterogeneity in the upper mantle is significant, and is organized into broad regions (continents and oceans)⁶⁰, shown by the wide correlation lengths. In the lower mantle, the correlation length diminishes to an approximately constant level until the base of the mantle is approached, when it increases again owing to the structure of D'' . There are no changes in correlation length in the lower mantle attributable to boundaries between elastically different material. Thus, whatever heterogeneity is present in the lower mantle, the observational seismic data only require it to be homogeneously distributed, without undergoing any significant scale reorganization with depth.

In summary, modern seismological studies show that the mantle is chemically heterogeneous at the 10-km length scale, but give no clear evidence of radial stratification other than at the boundary with the D'' region. Both tomographic and seismic array studies are consistent with deep subduction and hence with whole-mantle convection.

Heat flow and geochemistry

Current global heat loss is $44.2 \times 10^{12} \text{ W}$ (44.2 TW)⁶². This is made up of contributions from radioactive decay of K, U and Th and from secular cooling of the Earth. Measurements of the compositions of the most primitive, unfractionated, mantle peridotites indicate that

refractory lithophile elements such as Ca, Sc, Ti, the rare-earth elements (REE), U and Th are in chondritic ratio one to another in the silicate part of the Earth^{1,63}. This leads to concentrations of about 20 p.p.b. (parts per billion) U and 80 p.p.b. Th in bulk silicate Earth (that is, mantle + crust). Chemical analysis of igneous rocks indicates that their K/U ratios are generally about 12,500, irrespective of rock type or tectonic setting⁶⁴. This implies a K content of bulk silicate Earth of about 12,500 times the U content, or 250 p.p.m.

This inventory of radioactive elements is at present producing 20 TW, or about 45% of the total heat flow⁶⁵. But the mantle has a long thermal time constant, and advective removal of heat is linked to the vigour of plate-tectonic motions that vary over time, so that the heat loss measured now was generated by decay in the distant past. Although the age of the current heat-loss is, of course, model-dependent, we can make defensible estimates based on the apparent ages of isotopic heterogeneities in the mantle, such as those of modern **HIMU basalts**⁸. These are being generated now from melting of oceanic crust that was recycled from the surface into the deep interior 1–2 Gyr ago. Using the latter figures as a measure of the age of the heat reaching the surface yields a radioactive heat contribution of 24.9–32.9 TW (ref. 65). Parametrized convection models of the cooling of the Earth's interior yield secular cooling rates of 30–100 K Gyr⁻¹ (refs 66, 67). This loss of the Earth's heat energy produces heat fluxes between 5.9 and 20 TW, which are capable of making up the balance from the geochemically inferred radioactive heat production.

By mass-balancing the different parts of the silicate Earth to the overall bulk silicate composition¹, we can determine how much 'hidden' layering is necessary to make the whole. We will concentrate on the heat-producing elements, and begin by considering two major reservoirs whose concentrations of K, U and Th are reasonably well-known—the continental crust⁶⁸ and the mantle source of MORB⁶⁹. From the major-element compositions of MORB, it appears that these voluminous volcanic rocks are produced by about 10% melting of peridotite mantle^{70,71}. Assuming that K, U and Th act as **incompatible** elements, this means that the mantle

contains about 0.1 times the concentrations found in MORB. In that case the mantle source can only have about one-third the concentrations of K, U and Th (and other incompatible elements) postulated for the bulk silicate Earth¹. It is therefore 'depleted' mantle. As the continental crust is highly enriched in these incompatible elements, the simplest assumption is that enriched crust and depleted mantle are complementary, and formed by differentiation of bulk silicate Earth⁶⁹. The result (Table 1) indicates that there is insufficient enriched crust for the depleted peridotite mantle to comprise the whole mantle—it turns out to constitute only 50% of the whole mantle. The calculation can be repeated with isotope systems such as Sm–Nd with very similar results⁷². Mantle layering (Fig. 1) is the logical conclusion.

The simple layered model is, however, compromised by seismological evidence in favour of whole-mantle convection, or at least of deep subduction of slabs into the lower mantle and of heterogeneities introduced by subduction. That these heterogeneities are real and long-lasting is attested to by seismological evidence of lower-mantle 'scatterers', and by enriched ocean island basalts (OIBs) that contain isotopic evidence of contributions from old subducted oceanic crust (HIMU) and from continental crust and lithosphere (EM-1, EM-2)⁸. The current rate of subduction⁷³ of 3 km² per year means that approximately 18 km³ of oceanic crust—and, assuming 10% melting by mass, 140 km³ of highly depleted, 'sterile' peridotite—is being returned to the mantle each year. Rather more difficult to estimate is the amount of continental sediment that is returned to the mantle with the subducted oceanic plates. Recent estimates vary between 0.7 and 1.5 km³ per year⁷⁴. Therefore, assuming a constant rate of subduction for 4 Gyr, we find that the mantle contains 5% recycled oceanic crust, 45% recycled 'sterile' mantle and about 0.3% recycled continental material. If the higher rate of heat production in the past (3.6 times current values at 4 Gyr ago) was compensated by higher heat loss through formation of oceanic lithosphere, then it seems likely that almost the entire mantle has been through the cycle of partial melting, oceanic crust formation and sediment recycling. In that case a better view of the mantle today would be as a heterogeneous mixture of 'sterile' highly depleted mantle containing 'blobs' of oceanic and continental crust of different ages and sizes. Such 'blobs' are real, and recorded as chemical heterogeneities in studies of scattered seismic waves.

In Fig. 4 we show how a heterogeneous model of the mantle would fit with the chemical composition of the silicate Earth. Assuming that there is no primitive undepleted mantle, the silicate Earth consists of continental crust (0.6% by mass), 'sterile' mantle (> 45%), and recycled continental and oceanic material. As the depleted mantle produces basalt by about 10% partial melting^{70,71}, we can consider it a mixture of 10% recycled oceanic crust and 90% 'sterile' peridotite. There are two end-member approaches for mass-balancing K, U and Th in the silicate Earth that differ in their complexity. Either yields agreement with K, U and Th abundances, so we consider the differences they predict in other trace-element abundances to discriminate between them. One end-member assumes that average rates of oceanic crust production over the past 4 Gyr were twice the present rate, processing the whole mantle through the MORB production process, with 0.4% recycled continental crust material. The other end-member assumes a lower rate

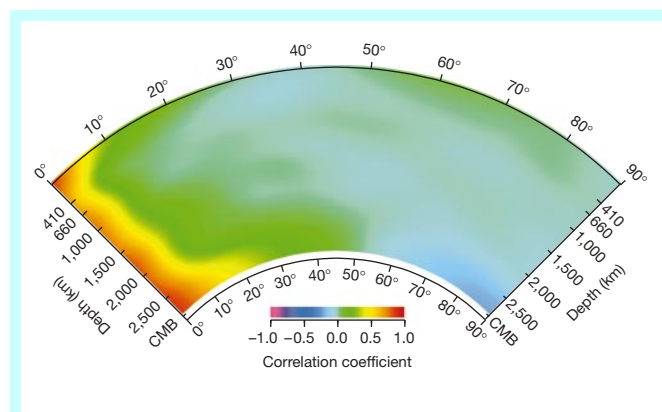


Figure 3 Angular correlation function of a shear-wave tomographic model³⁰ as a function of depth. This characterizes the similarity of wave speed of two points at the same depth in the mantle separated by some angular distance. Every point in the mantle is perfectly correlated with itself, resulting in the maximum correlation for points separated by zero distance (left side of figure). A drop of e^{-1} , or about 0.4, indicates a significant correlation decrease (the yellow band). Blue colours representing zero or negative correlations arise at large separations, and a degree-two pattern at the mantle's base is made evident by the maximum negative correlation at $\sim 90^\circ$. The correlation length is greatest in the upper mantle and in D". In the upper mantle, this probably corresponds to continent/ocean differences⁶⁰, while the structure in D" possibly reflects the pattern of heat flow out of the core into the mantle. In the lower mantle above D" the correlation length is approximately constant, indicating no reorganization of heterogeneity scale length in the mid–lower mantle. CMB, core–mantle boundary.

Table 1 Mass-balance calculation of the fraction of mantle depleted by continental crust formation

Element	Bulk silicate Earth (p.p.m.)	Continental crust (p.p.m.)	Depleted mantle (p.p.m.)	Fraction of mantle that is depleted
K	250	15,800	85	0.50
U	0.02	1.4	0.0065	0.54
Th	0.08	5.6	0.0164	0.46

of oceanic crust production, no recycled continental crust, and that the 16% remaining, yet-to-be-processed mantle mass fraction is excess—or ‘old’—material left over from an episode of early differentiation and oceanic crust extraction, as Hf isotope evidence suggests⁷⁵. In either case, modern MORB comes from a ‘marble-cake’⁵⁰ mixture of ‘sterile’ mantle and recycled oceanic crust. OIB has a similar origin, but the melting process must involve more of the recycled oceanic and continental crust in order to generate its higher incompatible trace-element content^{8,50}. We can test the two end-members by mass-balancing the bulk silicate Earth contents of elements such as Ti, Zr and La, which are geochemically similar to K, U and Th. The end-member with no ‘old’ recycled oceanic crust requires about 1% additional oceanic crust to mass-balance (Fig. 4). The end-member with ‘old’ excess crust overestimates the trace element abundances by an amount equivalent to 10% additional oceanic crust. Best overall agreement is achieved with a 75%/25% mixture of the end-member models (0.3% recycled continental material and 5% ‘excess’ recycled oceanic crust). In either, the recycled material in depleted mantle (>90% of the mantle) is

distributed as the small, kilometre-scale blobs discussed earlier. Recycled continental and ‘excess’ oceanic crust is largely isolated from the MORB reservoir, suggesting either a thin lower-mantle layer or that the lowermost mantle contains a high proportion of very long-lived blobs.

This simple model is a starting point for considering the mantle as a laterally heterogeneous, rather than a vertically stratified, system. In its simplicity, there are some geochemical aspects that it ignores. It ignores the role of delaminated continental crust, a possible component of EM-1 type OIB⁸, and the selective extraction and recycling of K and U by fluids in subduction zones⁷⁶. Nor do we consider any isotopic constraints on the necessity for primitive ‘untapped’ reservoirs somewhere in the Earth. Although both Sm–Nd and Rb–Sr isotope systematics can be fitted into the recycling scheme presented here, noble-gas isotopes should provide stronger constraints. The high ratio of primordial ³He to the radiogenic (from U, Th decay) ⁴He in some OIBs and the decoupling of heat from ⁴He flux are considered evidence of an undegassed deep reservoir^{7,10,84,85}. The bulk silicate Earth¹ has a K content which should have, over 4.5 Gyr, produced about twice as much ⁴⁰Ar as is observed in the atmosphere, crust and depleted mantle⁹, also implying that there is an isolated reservoir somewhere. These conclusions are, however, based on the assumption that the noble gases are highly incompatible and if this is not the case, they could be invalidated. For example, the ³He/⁴He ratio of all parts of the Earth is decreasing with time owing to production of ⁴He from U and Th. If, as seems possible, He is more compatible in mantle crystals than U and Th⁸⁶, then very old subducted mantle, which had undergone partial melting in the distant past, would have a relatively high ³He/U ratio. This region would retain a high ³He/⁴He ratio through geological time because of the low rate of ⁴He production. Some OIB might tap such a source rather than a primitive layer. Further complexities will clearly be required as the behaviour of the noble gases in OIB and MORB generation becomes better understood.

An integrated view of the mantle

Seismological and chemical evidence shows that heterogeneity is ubiquitous in the mantle, and that subduction is the key process producing it^{8,77}. This heterogeneity extends from top to bottom of the mantle under different guises. In the upper mantle, it takes the form predominantly of cold subducted lithospheric slabs. Their excess density carries them into the lower mantle where they either stagnate through thermal equilibration or sink to the core–mantle boundary. Simultaneously, the material mixes back into the mantle more effectively with time as thermal equilibration reduces rheological and density differences. Heterogeneity diminution by stretching at flow stagnation points yields a spectrum of sizes whose average diminishes with time but which still retains large fragments^{49,51}. D'' might also play a homogenizing role as a storage and entrainment site for oceanic crust^{44,78,79}. Rheological differences between the admixed material and the peridotitic mantle tend to inhibit and delay homogenization⁸⁰. In consequence, lower-mantle heterogeneity almost certainly exists at many scales, representing all stages in the mixing process. This is seismically expressed in the lower mantle as discrete lithospheric slabs, as 8-km-scale scatterers and in the unusual array of D'' properties.

How is this lower-mantle heterogeneity distributed and what proportion of it is likely to be seismically observable? To answer this we use the constraints of seismic scattering⁸¹. The volumetric fractional scattered power of a heterogeneity, $\langle \Phi^2 \rangle / VA^2$, depends on its volume a^3 , the average velocity perturbation v in the body, the scattering angle θ and wavenumber k , and decreases with distance r from the scatterer:

$$\left[\frac{\langle \Phi^2 \rangle}{VA^2} \right] = a^3 \left[\frac{k^4 v}{6\pi r^2} \right] \left[\frac{3 \cos \theta + 1 + 2 \cos^2 \theta}{1 + 4k^2 a^2 \sin^2(\theta/2)} \right]$$

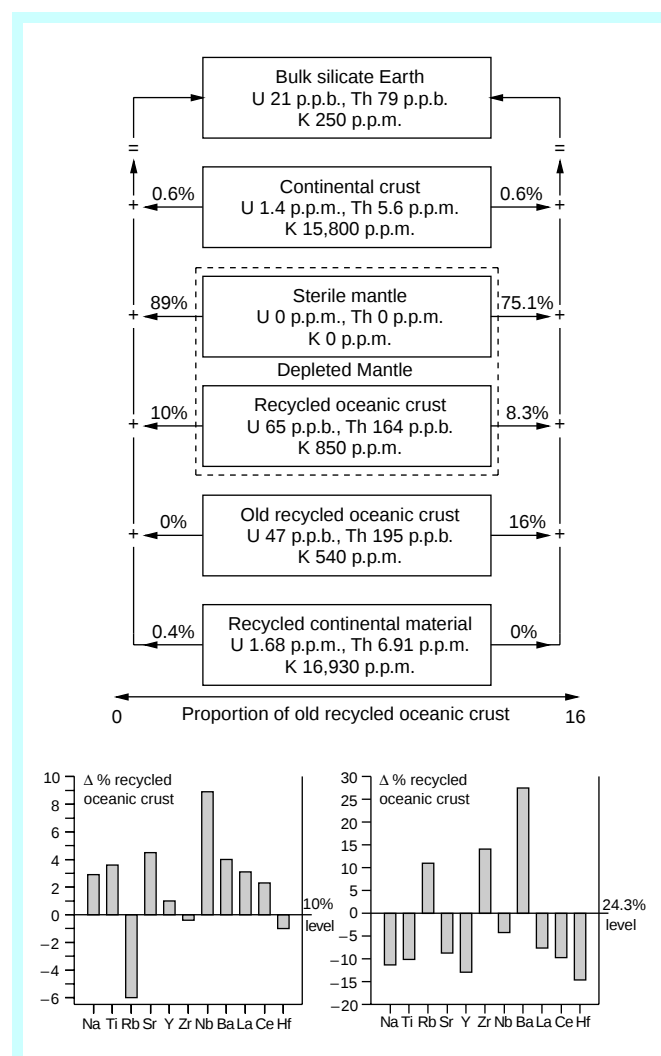


Figure 4 Approximate mass balance for a bulk silicate Earth made up of continental crust, ‘sterile’ mantle, recycled oceanic crust, old recycled oceanic crust, and recycled continental material. The proportions satisfy seismological, heat-flow and most trace-element geochemical constraints. Diagrams at bottom show misfit to trace-element abundances if the total amount of recycled oceanic crust is varied. The large Nb value is symptomatic of the ‘missing niobium problem’^{8,87}. Ba and Rb values are very sensitive to the amount of recycled continental crust and to unknown effects of subduction zone fluids.

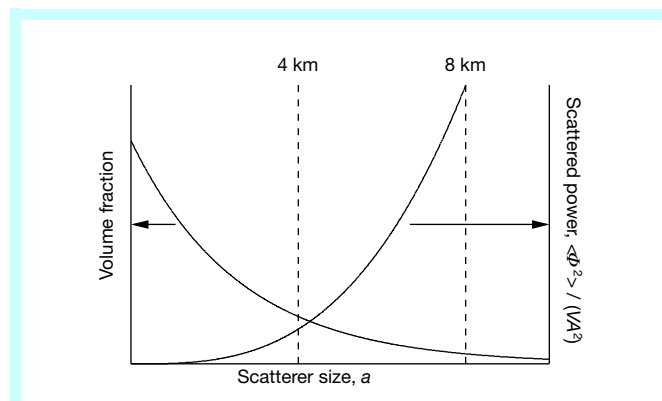


Figure 5 Visibility of mantle heterogeneity by seismic scattering. The abundance of mantle heterogeneities decreases with size exponentially (left axis), while the scattering power increases with the cube of the scatterer size (right axis). Scatterers smaller than ~ 4 km are not likely to be visible on account of the attenuation of high-frequency body waves in the upper mantle (wavenumber $k \propto$ frequency) that would be sensitive to them. Consequently, only the largest mantle heterogeneities yield observable scattering. With values of $\kappa = 2.25 \times 10^{-3} \text{ km}^{-1}$ and $f_0 = 2.04 \times 10^{-3}$ that satisfy the observational scattering constraints and the abundance of non-‘sterile’ mantle components, 93 vol.% of heterogeneity should be < 4 km in size, and hence invisible to seismic scattering.

Models of the convective stretching and dispersal of heterogeneities indicate that average size becomes smaller with time, with an exponential dependence of fractional volume f on size a (ref. 49): $f(a) = f_0 \exp(-\kappa a)$. The unknowns f_0 and κ can be constrained because the integral of f over heterogeneity sizes 0–8 km must yield the volume fraction of the mantle occupied by heterogeneities (around 11–16% from geochemical arguments). Also, the fraction at $a = 8$ km must agree with the number of seismically observed heterogeneities of that size. The latter is unknown, because much of the mantle has yet to be systematically explored for scatterers, but it should be of the order of the volume of subducted oceanic crust currently *en route* to the base of the mantle. This volume is about 0.2% of the mantle, assuming the present global rate of mass transfer into subduction zones and a vertical descent rate in the mantle of 30 mm yr^{-1} (refs 30, 48). Both constraints are satisfied by $\kappa = 2.25 \times 10^{-3} \text{ km}^{-1}$ and $f_0 = 2.04 \times 10^{-3}$, leading to the prediction that 93% (by volume) of the heterogeneity is smaller than 4 km, which would be invisible to short-period wave scattering (Fig. 5).

From a seismological viewpoint, the mantle therefore appears to exhibit an overlapping collection of vertically homogeneous regions that coexist with heterogeneity on a range of length scales (Fig. 6). However, most of the heterogeneity volume is represented by very small bodies (< 4 km). This homogeneous distribution of heterogeneity extends from the base of the mantle into the MORB melting region, 40–80 km below the surface, where it is destroyed by melting and melt extraction. The latter melting region acts as a near-surface ‘heterogeneity sink’—the larger heterogeneities can only be inferred from their chemical signatures. Despite having to traverse the region of homogenization, however, large peridotite bodies tectonically emplaced in the continental crust still exhibit a wide range of metre- and centimetre-scale heterogeneities^{82,83}.

Perspective and outstanding problems

Early interpretations of the seismic structure of the Earth, incorporating what was known about the compositions of meteorites, suggested a gradual change from silicate to iron-rich compositions with depth. This view was abandoned in favour of a layered mantle, once the nature of the core–mantle boundary was recognized, and finite strain theories became available to model changes in material elasticity with depth. The seismic discontinuities at 410 and 660 km

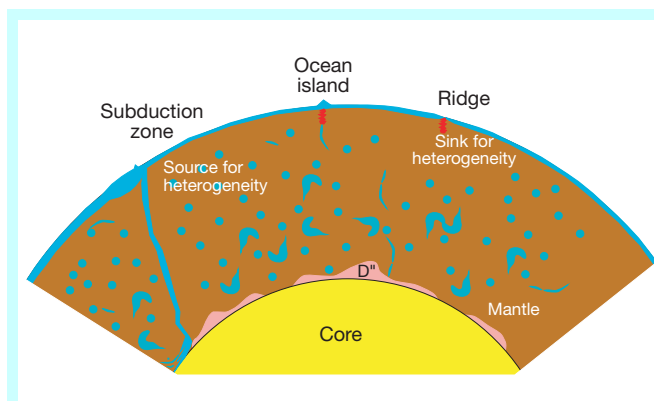


Figure 6 Sketch of proposed model of a chemically unstratified mantle. Subduction of oceanic lithosphere introduces heterogeneity into the mantle. Mixing by convective stirring of the mantle disaggregates the subducted lithosphere and minor continental material, producing isolated heterogeneities that scatter seismic energy but are too small to be observed tomographically. Melting at mid-ocean ridges and at ocean islands produces basalts and homogenizes the two types of mantle material, one enriched in incompatible elements and the other ‘sterile’.

depth initially appeared to be likely boundaries for the compositional layering. Recent studies of discontinuity topography and tomographic images of slab penetration into the lower mantle have, however, forced revision of this view. The latter results are most consistent with the discontinuities corresponding to phase transformations under isochemical conditions in a peridotite mantle. Stratification seems, therefore, to be an increasingly difficult position to defend. Lateral heterogeneity is, however, well established in the upper mantle, and has recently been shown to be present in the lower mantle. The principal issue is becoming how to interpret the heterogeneity seen at all depths in the mantle and to integrate it with the Earth’s chemical composition. We have presented here a simple geochemical model based on four components—‘sterile’ peridotite, recycled oceanic crust, and continental crust part of which is recycled—that is one way to explain most of the geophysical and geochemical observations.

With present techniques, seismic tomography, which averages over large volumes in the lower mantle, cannot resolve heterogeneities smaller than 400 km in size. Scattering, however, probes sub-wavelength scales in the mantle (~ 10 km), and the challenge will be to integrate this into whole-Earth studies. The aim now, guided by convective mixing models, will be to decipher the style and rate of mantle mixing, placing constraints on the residence times of the chemical components of the mantle in their respective reservoirs. The challenge to experimental and isotope geochemists is clearly to understand better the sources and sinks of rare gases in the Earth, and the mechanisms by which they are extracted from the mantle and deposited in the atmosphere. □

- McDonough, W. F. & Sun, S.-S. The composition of the Earth. *Chem. Geol.* **120**, 223–253 (1995).
- Newsom, H. E. in *Global Earth Physics: Handbook of Physical Constants* Vol. 1 (ed. Ahrens, T. J.) 159–189 (Reference Shelf Series, American Geophysical Union, Washington DC, 1995).
- Allègre, C. J., Manhès, G. & Gopel, C. The age of the Earth. *Geochim. Cosmochim. Acta* **59**, 1445–1456 (1995).
- Halliday, A. N., Rehkämper, M., Lee, D.-C. & Yi, W. Early evolution of the Earth and Moon: New constraints from Hf–W isotope geochemistry. *Earth Planet. Sci. Lett.* **142**, 75–89 (1996).
- Kellogg, L. H., Hager, B. H. & van der Hilst, R. D. Compositional stratification in the deep mantle. *Science* **283**, 1881–1884 (1999).
- Schilling, J.-G. Iceland mantle plume: geochemical evidence along Reykjanes Ridge. *Nature* **242**, 565–571 (1973).
- Allègre, C. J., Staudacher, T. & Sarda, P. Rare-gas systematics—formation of the atmosphere, evolution and structure of the Earth’s mantle. *Earth Planet. Sci. Lett.* **81**, 127–150 (1987).
- Hofmann, A. W. Mantle geochemistry: the message from oceanic volcanism. *Nature* **385**, 219–229 (1997).
- Allègre, C. J., Hofmann, A. W. & O’Nions, R. K. The argon constraints on mantle structure. *Geophys. Res. Lett.* **23**, 3555–3557 (1996).
- O’Nions, R. K. & Oxburgh, R. Heat and helium in the Earth. *Nature* **306**, 429–431 (1983).

11. Kennett, B. L. N., Engdahl, E. R. & Buland, R. Constraints on seismic velocities in the Earth from traveltimes. *Geophys. J. Int.* **122**, 108–124 (1995).
12. Lay, T., Williams, Q. & Garnero, E. J. The core–mantle boundary layer and deep Earth dynamics. *Nature* **392**, 461–468 (1998).
13. Bernal, J. D. Commentary. *Observatory* **59**, 265–269 (1936).
14. Ringwood, A. E. & Major, A. High-pressure transformations in pyroxenes. *Earth Planet. Sci. Lett.* **1**, 241–245 (1966).
15. Liu, L. The post-spinel phase of forsterite. *Nature* **262**, 770–772 (1976).
16. Ito, E. & Takahashi, E. Postspinel transformations in the system $\text{Mg}_2\text{SiO}_4\text{--Fe}_2\text{SiO}_4$ and some geophysical implications. *J. Geophys. Res.* **94**, 10637–10646 (1989).
17. Fei, Y., Saxena, S. & Navrotsky, A. Internally consistent thermodynamic data and equilibrium phase relations for compounds in the system MgO--SiO_2 at high pressure and temperature. *J. Geophys. Res.* **95**, 6915–6928 (1990).
18. Akaogi, M. & Akimoto, S. Pyroxene–garnet solid-solution equilibria in the systems $\text{Mg}_2\text{SiO}_4\text{--Mg}_2\text{Al}_2\text{Si}_2\text{O}_{12}$ and $\text{Fe}_2\text{SiO}_4\text{--Fe}_2\text{Al}_2\text{Si}_2\text{O}_{12}$ at high pressures and temperatures. *Phys. Earth Planet. Inter.* **15**, 90–106 (1977).
19. Ringwood, A. E. Phase transformations and their bearing on the constitution and dynamics of the mantle. *Geochim. Cosmochim. Acta* **55**, 2083–2110 (1991).
20. Wood, B. J. Phase transformations and partitioning relations in peridotite under lower mantle conditions. *Earth Planet. Sci. Lett.* **174**, 341–354 (2000).
21. Funamori, N., Yagi, T. & Utsumi, W. Thermoelastic properties of MgSiO_3 perovskite determined by in situ X-ray observations up to 30 GPa and 2000 K. *J. Geophys. Res.* **101**, 8257–8269 (1996).
22. Fiquet, G., Dewaele, A., Andraut, D., Kunz, M. & LeBihan, T. Thermoelastic properties and crystal structure of MgSiO_3 perovskite at lower mantle pressure and temperature conditions. *Geophys. Res. Lett.* **27**, 21–24 (2000).
23. Anderson, D. L. & Bass, J. D. Transition region of the Earth's upper mantle. *Nature* **320**, 321–328 (1986).
24. Jeanloz, R. & Knittle, E. Density and composition of the lower mantle. *Phil. Trans. R. Soc. Lond. A* **328**, 377–389 (1989).
25. Stixrude, L., Hemley, R. J., Fei, Y. & Mao, H. K. Thermoelasticity of silicate perovskite and magnesiowüstite and stratification of the Earth's mantle. *Science* **257**, 1099–1101 (1992).
26. Jeffreys, H. & Bullen, K. E. *Seismological Tables* (British Association Seismological Investigators Committee, London, 1940).
27. Shearer, P. M. Global mapping of upper-mantle reflectors from long-period SS precursors. *Geophys. J. Int.* **115**, 878–904 (1993).
28. Flanagan, M. P. & Shearer, P. M. Global mapping of topography on transition zone velocity discontinuities by stacking SS precursors. *J. Geophys. Res.* **103**, 2673–2692 (1998).
29. van der Hilst, R., Engdahl, E., Spakman, W. & Nolet, G. Tomographic imaging of subducted lithosphere below northwest Pacific island arcs. *Nature* **353**, 37–43 (1991).
30. Grand, S. P. Mantle shear structure beneath the Americas and surrounding oceans. *J. Geophys. Res.* **99**, 11591–11621 (1994).
31. Masters, G., Johnson, S., Laske, G. & Bolton, H. A shear velocity model of the mantle. *Phil. Trans. R. Soc. Lond. A* **354**, 1385–1410 (1996).
32. Vidale, J. E. & Benz, H. M. Upper mantle seismic discontinuities and the thermal structure of subduction zones. *Nature* **356**, 678–683 (1992).
33. Helffrich, G. R. & Bina, C. R. Frequency dependence of the visibility and depths of mantle seismic discontinuities. *Geophys. Res. Lett.* **21**, 2613–2616 (1994).
34. Stixrude, L. Structure and sharpness of phase transitions and mantle discontinuities. *J. Geophys. Res.* **102**, 14835–14852 (1997).
35. McKenzie, D. P. Speculations on the consequences and causes of plate motions. *Geophys. J. R. Astron. Soc.* **18**, 1–32 (1969).
36. Bina, C. R. & Helffrich, G. Phase-transition Clapeyron slopes and transition zone seismic discontinuity topography. *J. Geophys. Res.* **99**, 15853–15860 (1994).
37. Collier, J. & Helffrich, G. Topography of the “410” and “660” km seismic discontinuities in the Izu-Bonin subduction zone. *Geophys. Res. Lett.* **24**, 1535–1538 (1997).
38. Castle, J. C. & Creager, K. C. Seismic evidence against a mantle chemical discontinuity near 660 km depth beneath Izu-Bonin. *Geophys. Res. Lett.* **24**, 241–244 (1997).
39. Wicks, C. W. & Richards, M. A. A detailed map of the 660-kilometer discontinuity beneath the Izu-Bonin subduction zone. *Science* **261**, 1424–1427 (1993).
40. Christensen, U. & Yuen, D. The interaction of a subducting lithospheric slab with a chemical or phase boundary. *J. Geophys. Res.* **89**, 4389–4402 (1984).
41. Kincaid, C. & Olson, P. An experimental study of subduction and slab migration. *J. Geophys. Res.* **92**, 13822–13840 (1987).
42. Bijwaard, H., Spakman, W. & Engdahl, E. R. Closing the gap between regional and global travel time tomography. *J. Geophys. Res.* **103**, 30055–30078 (1998).
43. Davies, G. & Richards, M. Mantle convection. *J. Geol.* **100**, 151–206 (1992).
44. Christensen, U. R. & Hofmann, A. W. Segregation of subducted oceanic crust in the convecting mantle. *J. Geophys. Res.* **99**, 19867–19884 (1994).
45. Hedlin, M. A. H., Shearer, P. M. & Earle, P. S. Seismic evidence for small-scale heterogeneity throughout the Earth's mantle. *Nature* **387**, 145–150 (1997).
46. Vidale, J. E. & Hedlin, M. A. H. Evidence for partial melt at the core–mantle boundary north of Tonga from the strong scattering of seismic waves. *Nature* **391**, 682–685 (1998).
47. Castle, J. C. & Creager, K. C. Topography of the 660-km seismic discontinuity beneath Izu-Bonin: Implications for tectonic history and slab deformation. *J. Geophys. Res.* **103**, 12511–12527 (1998).
48. Kaneshima, S. & Helffrich, G. Dipping low-velocity layer in the mid-lower mantle: Evidence for geochemical heterogeneity. *Science* **283**, 1888–1891 (1999).
49. Olson, P., Yuen, D. A. & Balsiger, D. Convective mixing and the fine structure of mantle heterogeneity. *Phys. Earth Planet. Inter.* **36**, 291–304 (1984).
50. Allègre, C. J. & Turcotte, D. L. Implications of a two-component marble-cake mantle. *Nature* **323**, 123–127 (1986).
51. Gurnis, M. & Davies, G. F. Mixing in numerical models of mantle convection incorporating plate kinematics. *J. Geophys. Res.* **91**, 6375–6395 (1986).
52. Chai, M., Brown, J. M. & Slutsky, L. J. Thermal diffusivity of mantle minerals. *Phys. Chem. Minerals* **23**, 470–475 (1996).
53. Weber, M. & Davis, J. P. Evidence of a laterally variable lower mantle structure from P- and S-waves. *Geophys. J. Int.* **102**, 231–255 (1990).
54. Earle, P. S. & Shearer, P. M. Observations of PKKP precursors used to estimate small-scale topography on the core–mantle boundary. *Science* **277**, 667–670 (1997).
55. Yamada, A. & Nakanishi, I. Short-wavelength lateral variation of a D⁺ P-wave reflector beneath the southwestern Pacific. *Geophys. Res. Lett.* **25**, 4545–4548 (1998).
56. Thomas, C., Weber, M., Wicks, C. W. & Scherbaum, F. Small scatterers in the lower mantle observed at German broadband arrays. *J. Geophys. Res.* **104**, 15073–15088 (1999).
57. Revenaugh, J. & Meyer, R. Seismic evidence of partial melt within a possibly ubiquitous low-velocity layer at the base of the mantle. *Science* **277**, 670–673 (1997).
58. Garnero, E. J. & Helmberger, D. V. Further structural constraints and uncertainties of a thin laterally varying ultralow-velocity layer at the base of the mantle. *J. Geophys. Res.* **103**, 12495–12509 (1998).
59. Kendall, J. M. & Silver, P. G. Constraints from seismic anisotropy on the nature of the lowermost mantle. *Nature* **381**, 409–412 (1996).
60. Su, W.-J. & Dziewonski, A. M. Simultaneous inversion for 3-D variations in shear and bulk velocity in the mantle. *Phys. Earth Planet. Inter.* **100**, 135–156 (1997).
61. Grand, S. P., van der Hilst, R. D. & Widiyantoro, S. Global seismic tomography: A snapshot of convection in the Earth. *GSA Today* **7**, 1–7 (1997).
62. Pollack, H. N., Hurter, S. J. & Johnston, J. R. Heat flow from the Earth's interior—analysis of the global data set. *Rev. Geophys.* **31**, 267–280 (1993).
63. Hart, S. R. & Zindler, A. In search of a bulk-Earth composition. *Chem. Geol.* **57**, 247–267 (1986).
64. Jochum, K. P., Hofmann, A. W., Ito, E., Seufert, H. M. & White, W. M. K, U and Th in mid-ocean ridge basalt glasses and heat production, K/U and K/Rb in the mantle. *Nature* **306**, 431–436 (1986).
65. Van Schmus, W. R. in *Global Earth Physics: Handbook of Physical Constants* Vol. 1 (ed. Ahrens, T. J.) 283–291 (Reference Shelf Series, American Geophysical Union, Washington DC, 1995).
66. Jackson, M. J. & Pollack, H. N. On the sensitivity of parameterized convection to the rate of decay of internal heat sources. *J. Geophys. Res.* **89**, 10103–10108 (1984).
67. Christensen, U. Thermal evolution models for the Earth. *J. Geophys. Res.* **90**, 2995–3007 (1985).
68. Rudnick, R. L. Making continental crust. *Nature* **378**, 571–578 (1995).
69. Hofmann, A. W. Chemical differentiation of the Earth: the relationship between mantle, continental crust and oceanic crust. *Earth Planet. Sci. Lett.* **90**, 297–314 (1988).
70. Kinzler, R. J. & Grove, T. L. Primary magmas of midocean ridge basalts. 2. Applications. *J. Geophys. Res.* **97**, 6907–6926 (1992).
71. Klein, E. M. & Langmuir, C. H. Global correlations of ocean ridge basalt chemistry with axial depth and crustal thickness. *J. Geophys. Res.* **92**, 8089–8115 (1987).
72. Allègre, C. J. Chemical geodynamics. *Tectonophysics* **81**, 109–132 (1982).
73. Chase, C. G. The n-plate problem of plate tectonics. *Geophys. J. R. Astron. Soc.* **29**, 117–122 (1972).
74. Plank, T. & Langmuir, C. H. The geochemical composition of subducting sediment and its consequences for the crust and mantle. *Chem. Geol.* **145**, 325–394 (1998).
75. Blichert-Toft, J. & Albarède, F. The Lu–Hf isotope geochemistry of chondrites and the evolution of the mantle–crust system. *Earth Planet. Sci. Lett.* **149**, 243–258 (1997).
76. Elliott, T., Zindler, A. & Bourdon, B. Exploring the kappa conundrum: the role of recycling in the lead isotope evolution of the mantle. *Earth Planet. Sci. Lett.* **169**, 129–145 (1999).
77. Davies, G. F. Earth's neodymium budget and structure and evolution of the mantle. *Nature* **290**, 208–213 (1981).
78. Hofmann, A. W. & White, W. M. Mantle plumes from ancient oceanic crust. *Earth Planet. Sci. Lett.* **57**, 421–436 (1982).
79. Sleep, N. H. Gradual entrainment of a chemical layer at the base of the mantle by overlying convection. *Geophys. J. Int.* **95**, 437–448 (1988).
80. Metcalfe, G., Bina, C. R. & Ottino, J. M. Kinematic considerations for mantle mixing. *Geophys. Res. Lett.* **22**, 743–746 (1995).
81. Chernov, L. A. *Wave Propagation in a Random Medium* (Dover, New York, 1960).
82. Kornprobst, J. Le massif ultrabasique des Beni Bouchera (Rif Interne, Maroc): étude des péridotites de haute température et de haute pression, et des pyroxénites, à grenat ou sans grenat, qui leur sont associées. *Contrib. Mineral. Petrol.* **23**, 283–322 (1969).
83. Nicolas, A., Bouchez, J. L. & Boudier, F. Interprétation cinématique des déformations plastiques dans le massif de l'herzolite de Lanzo (Alpes Piémontaises)—comparaison avec d'autres massifs. *Tectonophysics* **14**, 143–171 (1972).
84. Porcelli, D. & Wasserburg, G. J. Mass-transfer of helium, neon, argon, and xenon through a steady-state upper-mantle. *Geochim. Cosmochim. Acta* **59**, 4921–4937 (1995).
85. O'Nions, R. K. & Tolstikhin, I. N. Limits on the mass flux between lower and upper mantle and stability of layering. *Earth Planet. Sci. Lett.* **139**, 213–222 (1996).
86. Broadhurst, C. L., Drake, M. J., Hagee, B. E. & Bernatowicz, T. J. Solubility and partitioning of Ar in anorthite, diopside, forsterite, spinel, and synthetic basaltic liquids. *Geochim. Cosmochim. Acta* **54**, 299–309 (1990).
87. Rudnick, R. L., Barth, M., Horn, I. & McDonough, W. F. Rutile-bearing refractory eclogites: Missing link between continents and depleted mantle. *Science* **287**, 278–281 (2000).

Correspondence and requests for materials should be addressed to G.R.H.
(e-mail: george@geology.bristol.ac.uk).

Ground-based observation of emission lines from the corona of a red-dwarf star

J. H. M. M. Schmitt & R. Wichmann

Universität Hamburg, Hamburger Sternwarte, 21029 Hamburg, Gojenbergsweg 112, Germany

All 'solar-like' stars¹ are surrounded by coronae², which contain magnetically confined plasma at temperatures above 10^6 K. (Until now, only the Sun's corona could be observed in the optical—as a shimmering envelope during a total solar eclipse.) As the underlying stellar 'surfaces'—the photospheres—are much cooler, some non-radiative process must be responsible for heating the coronae. The heating mechanism is generally thought to be magnetic in origin, but is not yet understood even for the case of the Sun. Ultraviolet emission lines first led to the discovery of the enormous temperature of the Sun's corona^{3,4}, but thermal emission from the coronae of other stars has hitherto been detectable only from space, at X-ray wavelengths. Here we report the detection of emission from highly ionized iron (Fe XIII at 3,388.1 Å) in the corona of the red-dwarf star CN Leonis, using a ground-based telescope. The X-ray flux inferred from our data is consistent with previously measured X-ray fluxes, and the non-thermal line width of 18.4 km s^{-1} indicates great similarities between solar and stellar coronal heating mechanisms. The accessibility and spectral resolution (45,000) of the ground-based instrument are much better than those of X-ray satellites, so a new window to the study of stellar coronae has been opened.

Coronal plasma at temperatures above 10^6 K emits the bulk of its radiative energy in the X-ray range. Therefore X-ray emission is the most natural way to detect and diagnose coronal emissions. Non-thermal gyrosynchrotron emission has been detected from a substantial number of stars⁵ in the radio range, but the radio stars are quite different from the Sun, which, if placed at stellar distances, would not be detectable as a radio source. The advent of soft X-ray imaging on board the Einstein (1978–81) and Rosat (1990–98) satellites has led to the discovery of X-ray emission from many thousands of solar-like stars^{6,7}. Small fractions of the total coronal energy losses are, however, emitted in the extreme-ultraviolet and far-ultraviolet bands⁸—which can only be accessed from space—as well as in the optical band. The optical band is interesting because ground-based telescopes rather than satellites can be used for data collection; observing from the ground is far more economical than observing from space.

At optical wavelengths, the corona of the Sun can be studied most conveniently during a total solar eclipse or with the help of a coronagraph. The coronal emission observed at optical wavelengths is mostly photospheric light scattered off free electrons or interplanetary dust particles (termed 'K and F corona'). A small fraction (~1%) of the optical solar coronal emission comes as line emission (termed 'L corona'). The three strongest of these optical coronal emission lines have wavelengths of 3,388 Å (Fe XIII), 5,303 Å (Fe XIV) and 6,374 Å (Fe X). These lines arise from transitions within the ions' ground configuration, and thus are not 'allowed' electric dipole transitions but rather 'forbidden' magnetic dipole or electric quadrupole transitions. The identification of these 'forbidden' lines had led to the realization that the solar corona is a plasma much thinner and hotter than the photosphere^{3,4}, thus raising the 'coronal heating problem', a puzzle still unsolved today.

In a stellar context, it is impossible to block the photospheric stellar light with the Moon or a coronagraph. The weak coronal

emission has to be detected against the full photospheric background, and thus one can only hope to detect coronal line emission ('L corona'). Red-dwarf stars are much fainter than the Sun, yet their X-ray luminosities are as large as—or even larger than—that of the Sun. Therefore red-dwarf stars with their low fluxes in the blue and ultraviolet ranges are the obvious candidates to search for forbidden coronal line emission, but all previous attempts to detect such lines have failed because of a lack of ultraviolet sensitivity⁹.

We have observed the red dwarf CN Leo using the Ultraviolet–Visual Echelle Spectrograph (UVES)¹⁰ mounted on ESO's Kueyen telescope. The exposure time was 3,120 seconds, and the covered spectral range was 3,100–3,800 Å with a spectral resolution of ~45,000. CN Leo is a nearby (distance $d = 2.4$ pc) optically faint star (luminosity $L_{\text{CN Leo}} = 10^{-3} L_{\odot}$, where $L_{\odot}$ is the solar luminosity) of spectral type M6V. The star shows H_{α} emission and frequent optical flares¹¹. It is a strong X-ray source, detected with both the Einstein¹² and Rosat satellites¹³—its X-ray output is comparable to, and often exceeds, that of the Sun.

Our observations were geared towards a detection of the forbidden Fe XIII line at 3,388.1 Å. The recorded UVES spectrum of CN Leo in the range 3,370–3,395 Å, shifted with a velocity of $+5.54 \text{ km s}^{-1}$ to the star's rest frame, is shown in Fig. 1a. This part of the spectrum is dominated by chromospheric Ti II emission lines. The measured transitions and derived line parameters are

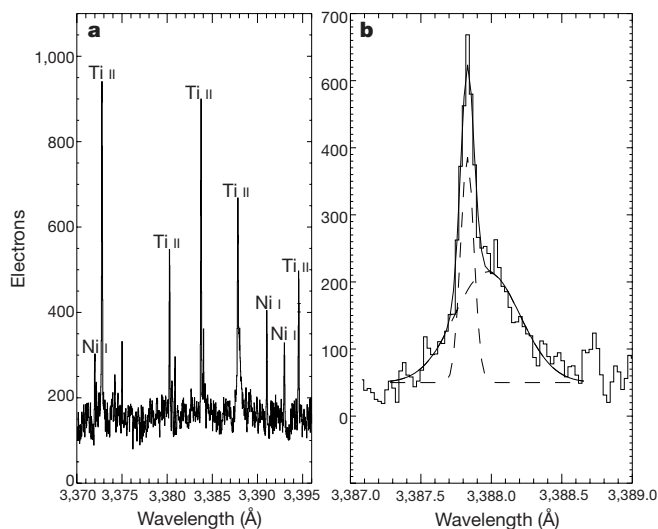


Figure 1 The ultraviolet spectrum of CN Leo from 3,370 Å to 3,395 Å. **a**, The recorded UVES spectrum of CN Leo in the range 3,370–3,395 Å. Here, as in Fig. 2, the recorded CCD electrons, obtained by multiplying the ADU (analog-to-digital units) output by 0.6. The strong emission lines (marked by Ti II) are all due to chromospheric Ti II lines. They correspond to transitions from the configuration $3d^24p$ (4G) with total angular momentum levels at $J = 5/2, 7/2, 9/2$ and $11/2$ into the ground state configuration $3d^24s$ (4F) with levels $J = 3/2, 5/2, 7/2$ and $9/2$; the ground state has $J = 3/2$. A large number of transitions satisfy the selection rule ($\Delta J = 0, \pm 1$) of allowed transitions. We note that transitions between the same configurations are expected to have the same line shapes; this is the case except for the $\text{Ti II } ^4F_{7/2} - ^4G_{7/2}$ transition at 3,387.85 Å, which is clearly different. **b**, Plot of the recorded UVES spectrum between 3,387 and 3,389 Å (solid line), together with a simple model fit (dashed curves). The narrow line component (due to Ti II) was fitted with a gaussian line with a width kept fixed at the line widths of the other Ti II lines (compare Table 1). The broad line profile was fitted with a gaussian line allowing amplitude, central wavelength and line width to vary freely; the resulting fit parameters are listed in Table 1. The additional broad spectral component is shifted by 0.13 Å with respect to the narrow Ti II line at 3,387.85 Å and has a gaussian FWHM of 0.52 Å as compared to a dispersion of 0.09 Å for the narrow line components. The central wavelength of the broad line with barycentric corrections applied agrees very well with the expected wavelength of the forbidden iron line at 3,388.1 Å.

summarized in Table 1. In Fig. 1a, the $\text{Ti II } ^4\text{F}_{7/2} - ^4\text{G}_{7/2}$ transition at 3,387.85 Å appears much broader than the other Ti II lines. In Fig. 1b we show the enlarged (and shifted) UVES spectrum between 3,387 and 3,389 Å, together with a simple model fit consisting of a narrow and broad line. In the fit, the full-width at half-maximum (FWHM) of the narrow line was kept constant at a value of 0.092 Å, consistent with the widths of all the other Ti II emission lines. The central wavelength of the broad line component agrees very well with the expected wavelength of the forbidden Fe XIII line, and therefore we identify this broad component with the Fe XIII 3,388.1 Å line. This interpretation is strongly supported by the observed line width. Correcting the observed linewidth σ_λ by the instrumental dispersion and converting to temperature with the formula $T = (\sigma_\lambda c / \lambda)^2 (M/k)$ (where M is the atomic mass of iron, and k the Boltzmann constant), we find a temperature of 2.75×10^6 K. We emphasize that this temperature refers to the ion kinetic temperature, and not the ionization temperature. The ion kinetic temperature could be lower if there are turbulent velocity fields. If we assume—as is usually done—that the ion kinetic temperature is very close to the peak formation temperature¹⁴ of Fe XIII of 1.6×10^6 K, we require an additional turbulent broadening of $v_{\text{turb}} = 18.4 \text{ km s}^{-1}$ to explain the observed width of the Fe XIII 3,388.1 Å line.

Conversion of the measured line count into an energy flux yields a value of $f_{3388} = 3 \times 10^{-16} \text{ erg cm}^{-2} \text{ s}^{-1}$. We estimate the accuracy of this value to be about 50%, because no precise photometric calibrations were performed. Given the distance of 2.4 pc towards CN Leo, this results in a total line luminosity L_{3388} of $2.1 \times 10^{23} \text{ erg s}^{-1}$ or a line photon emissivity of N_{3388} of

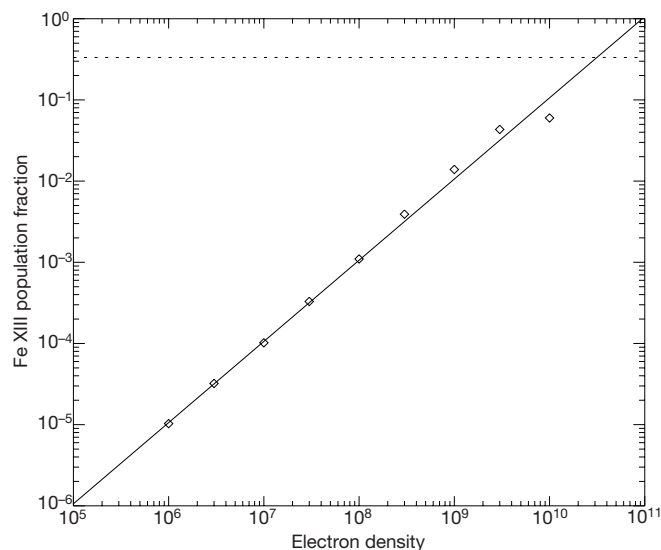


Figure 2 The volume emission measure of CN Leo from the Fe XIII line at 3,388 Å. Data points, the population fraction f of the $^1\text{D}_2$ level as a function of electron density N_e according to the calculations of ref. 15; the calculations were carried out for a temperature of 2.5×10^6 and a dilution factor $W = 0.1$ for the radiation field. The solid line shows the curve $f = 1.06 \times 10^{-11} N_e$, indicating that f and N_e are directly proportional up to electron densities of about 10^{11} cm^{-3} . At high electron densities, when collisional de-excitation dominates all ground configuration levels, the population fraction becomes asymptotically $f = 1/3$, indicated by the horizontal dotted line. Knowledge of f allows the computation of the volume emission measure $N_e^2 V$ from the measured photon luminosity N_{3388} as follows. With $A_{3388} = 78.4 \text{ s}^{-1}$ denoting the Einstein coefficient¹⁵ of the considered transition, V the coronal volume, $\xi_{\text{Fe XIII}}$ the ionization fraction of Fe XIII , taken as 0.25 for $T = 1.6 \times 10^6$ K and 0.01 for $T = 2.75 \times 10^6$ K, $A_{\text{Fe}} (= 2.6 \times 10^{-5})$ the iron abundance with respect to hydrogen, N_e the electron density and ξ_e the ratio between proton and electron densities, we have $N_{3388} = A_{3388} V f_{\text{Fe XIII}} \xi_{\text{Fe XIII}} \xi_e N_e = 1.06 \times 10^{-11} A_{3388} A_{\text{Fe}} \xi_{\text{Fe XIII}} \xi_e N_e^2 V$ and thus the emission measure $N_e^2 V$ can be inferred from the observations.

$3.6 \times 10^{34} \text{ photons s}^{-1}$. From N_{3388} , the volume emission measure $N_e^2 V$ can be calculated (see Fig. 2 legend; data taken from ref. 15). Performing such a calculation under the assumptions (1) $T = 1.6 \times 10^6$ K, $v_{\text{turb}} = 18.4 \text{ km s}^{-1}$ and (2) $T = 2.75 \times 10^6$ K, $v_{\text{turb}} = 0 \text{ km s}^{-1}$, we find total volume emission measures ($N_e^2 V$) of $7.8 \times 10^{48} \text{ cm}^{-3}$ and $1.9 \times 10^{50} \text{ cm}^{-3}$ (see Fig. 2), which leads to broad-band X-ray luminosities of about $3.1 \times 10^{26} \text{ erg s}^{-1}$ and $7.8 \times 10^{27} \text{ erg s}^{-1}$ for cases (1) and (2), respectively. These numbers can be compared to previous X-ray observations. The Imaging Proportional Counter (IPC) on board the Einstein satellite observed CN Leo on two occasions (21 May and 5 December 1979)¹⁶, giving X-ray luminosities of $L_X \approx 8.5 \times 10^{26} \text{ erg s}^{-1}$ and $L_X \approx 3.6 \times 10^{26} \text{ erg s}^{-1}$, respectively. With the Rosat Position Sensitive Proportional Counter (PSPC), CN Leo was observed on 4 December 1993. The count rate continuously declined during these observations, indicating the presence of a strong flare. At the end of the PSPC observations the equivalent X-ray luminosity was $\sim 1.0 \times 10^{27} \text{ erg s}^{-1}$, a value which needs to be considered as an upper limit to the actual quiescent X-ray luminosity. We therefore conclude that case (1) fits much better to the existing X-ray data than case (2), which would result in an X-ray luminosity about one order of magnitude larger than ever observed. Of course, this agreement may be fortuitous because the X-ray data were taken many years before our UVES observations.

The volume emission measures derived from the Fe XIII 3,388 Å line strengths favour an interpretation of turbulent broadening up to 18.4 km s^{-1} ; assuming no turbulent broadening leads to unrealistic X-ray luminosities. This conclusion is also supported by observations at high spectral resolution of solar coronal emission lines. The forbidden 1,242.03 Å line from Fe XIII (formed at 1.4×10^6 K as compared to 1.6×10^6 K for Fe XIII) and the non-forbidden lines (609 and 625 Å) of Mg X (formed at 1.0×10^6 K) have non-thermal widths of respectively $16.5 \pm 7.0 \text{ km s}^{-1}$ and $23.2 \pm 5.0 \text{ km s}^{-1}$, both for active and quiescent coronal regions in the Sun^{17,18}. Thus the solar values for turbulent broadening in Fe XII agree exactly with those measured in Fe XIII for CN Leo. This suggests that coronal dynamics and coronal heating are very similar in the Sun and in CN Leo.

Our Kueyen/UVES discovery of the Fe XIII 3,388.1 Å line can be placed in a historical context. The solar corona was identified as a hot, tenuous gas when emission lines were identified with transitions in highly ionized atoms about 60 years ago^{6,7}. Space-based X-ray observations have shown the presence of coronae around other stars, but it has hitherto been technically impossible to detect the same emission lines that were first observed from the Sun. In this sense, the present work closes the historical loop, by showing that it is possible to study the physical conditions in stellar coronae using ground-based observations.

Ground-based instrumentation outperforms the current (and

Table 1 Chromospheric and coronal emission lines in CN Leo

Line identification	Wavelength, obs. (Å)	Wavelength, exp. (Å)	Line width FWHM (Å)	Line count
$\text{Ti II } ^4\text{F}_{5/2} - ^4\text{G}_{7/2}$	3,372.79	3,372.80	0.106 ± 0.009	5,692
$\text{Ti II } ^4\text{F}_{9/2} - ^4\text{G}_{9/2}$	3,380.27	3,380.28	0.092 ± 0.009	2,687
$\text{Ti II } ^4\text{F}_{3/2} - ^4\text{G}_{5/2}$	3,383.75	3,383.77	0.092 ± 0.008	4,195
$\text{Ti II } ^4\text{F}_{5/2} - ^4\text{G}_{5/2}$	3,394.56	3,394.58	0.106 ± 0.012	2,700
$\text{Ti II } ^4\text{F}_{7/2} - ^4\text{G}_{7/2}$	3,387.83	3,387.85	0.094 (fixed)	2,315
$\text{Fe XIII } ^3\text{P}_2 - ^1\text{D}_2$	3,387.96	3,388.1	0.535 ± 0.059	6,773

Summary of flux and width measurements for the lines displayed in Fig. 1a; given are the laboratory and observed wavelengths (corrected for the radial velocity of CN Leo, 5.54 km s^{-1}), the measured line strengths (in terms of analog to digital units), the gaussian FWHM values as well as the transition identification. The line of interest is the Fe XIII transition. The $3s^2 3p^2$ ground configuration of Fe XIII splits up into the states $^3\text{P}_{2,1,0}$, $^1\text{D}_1$ and $^1\text{S}_0$, the ground state being $^3\text{P}_0$. The transitions $^3\text{P}_2 - ^3\text{P}_1$ and $^3\text{P}_1 - ^3\text{P}_0$ lie in the infrared at 10,796 and 10,747 Å, respectively. The $^1\text{D}_2 - ^3\text{P}_2$ transition gives rise to the observed 3,388.1 Å line.

next) generation of X-ray spectrometers by more than an order of magnitude in spectral resolution. Such improved capability has enabled us to obtain a ground-based measurement of the kinetic temperature and turbulent velocity of a coronal ion. The observed good agreement between the observed and expected wavelengths of the Ti II and Fe XIII lines indicates that the chromosphere and the corona of CN Leo are not in relative motion with respect to each other as expected for a quiescent corona. CN Leo is a flare star—albeit not flaring (at least in a major way) during our UVES observations. During a flare mass motions must occur, possibly with speeds of up to a few hundred kilometres per second. A line-of-sight velocity of 100 km s^{-1} would lead to a $1.13 \text{ \AA}$ shift of the Fe XIII line, which should be easily detectable with UVES. Thus it should be possible to directly diagnose the coronal dynamics of flares on CN Leo and other stars.

It is a challenge to extend coronal line observations to a larger sample of stars, and to stars of earlier spectral type. With such samples, ground-based searches for stellar coronal activity cycles—similar to those on the Sun—could be carried out. For sufficiently rapid rotators, Doppler tomography could be used to study stellar coronal structure. With yet deeper exposures, it should be possible to detect weaker coronal lines (for example, Ca XII at $3,329 \text{ \AA}$), and interesting plasma diagnostics would be obtained by extending the coronal line observations to $5,403 \text{ \AA}$ (Fe XIV), $6,374 \text{ \AA}$ (Fe X), and $10,796$ and $10,747 \text{ \AA}$ (Fe XIII). As the photospheric background is much stronger at longer wavelengths, this will be a challenge both for observations and atmospheric modelling. $\square$

Received 2 May; accepted 18 May 2001.

1. Linsky, J. L. Non radiative activity across the H-R diagram—Which types of stars are solar-like? *Sol. Phys.* **100**, 333–362 (1985).
2. Schmitt, J. H. M. M. Coronae on solar-like stars. *Astron. Astrophys.* **318**, 215–230 (1997).
3. Grottian, W. Zur Frage der Deutung der Linien im Spektrum der Sonnenkorona. *Naturwissenschaften* **27**, 214 (1939).
4. Edlén, B. Die Deutung der Emissionslinien im Spektrum der Sonnenkorona. *Z. Astrophys.* **22**, 30–62 (1942).
5. Taylor, A. R. & Paredes, J. M. (eds) *Radio Emission from the Stars and the Sun* (ASP Conf. Ser. 93, Astronomical Society of the Pacific, San Francisco, 1996).
6. Vaiana, G. S. *et al.* Results from an extensive Einstein stellar survey. *Astrophys. J.* **244**, 163–182 (1981).
7. Guilloit, P. *et al.* The stellar content of soft X-ray surveys. II. Cross-correlation of the ROSAT All-Sky Survey with the Tycho and Hipparcos catalogs. *Astron. Astrophys.* **351**, 1003–1015 (1999).
8. Linsky, J. L., Wood, B. E., Brown, A. & Osten, R. A. Dissecting Capella's corona: GHRS spectra of the Fe XXI 1354 and He II λ 1640 lines from each of the capella stars. *Astrophys. J.* **492**, 767–777 (1998).
9. Wallerstein, G., Böhm, K. H. & Oke, J. B. in *Extreme Ultraviolet Astronomy* (eds Malina, R. F. & Bowyer, S.) 113 (Pergamon, New York, 1991).
10. D'Odorico, S., Kaper, L. & Kaufer, A. *VLT UVES Manual* at (<http://www.eso.org/instruments/uves/userman/>) (2000).
11. Robinson, R. D., Carpenter, K. G., Percival, J. W. & Bookbinder, J. A. A search for microflaring activity on dMe flare stars. I. Observations of the dM8e star CN Leonis. *Astrophys. J.* **451**, 795–805 (1995).
12. Barbera, M., Micela, G., Sciortino, S., Harnden, F. R. Jr & Rosner, R. X-ray emission at the low-mass end—Results from an extensive Einstein Observatory survey. *Astrophys. J.* **414**, 846–866 (1993).
13. Schmitt, J. H. M. M., Fleming, T. A. & Giampapa, M. S. The X-ray view of the low-mass stars in the solar neighborhood. *Astrophys. J.* **450**, 392–400 (1995).
14. Arnaud, M. & Raymond, J. Iron ionization and recombination rates and ionization equilibrium. *Astrophys. J.* **398**, 394–406 (1992).
15. Flower, D. R. & Pineau Des Forets, G. Excitation of the Fe XIII spectrum in the solar corona. *Astron. Astrophys.* **24**, 181–192 (1973).
16. Harris, D. E. *et al.* *The Einstein Observatory Catalog of IPC X-ray Sources*, Vol. 4 (Smithsonian Astrophysical Observatory, Cambridge, Massachusetts, 1990).
17. Cheng, C., Doschek, G. A. & Feldman, U. The dynamical properties of the solar corona from intensities and emission line widths of EUV forbidden lines. *Astrophys. J.* **227**, 1037–1046 (1979).
18. Chae, J., Schühle, U. & Lemaire, P. SUMER measurements of non thermal motions: Constraints on coronal heating mechanisms. *Astrophys. J.* **505**, 957–973 (1998).

Acknowledgements

The data reported in this paper were collected at ESO's Paranal Observatory. We gratefully acknowledge the support of the ESO staff. We also thank D. Flower for discussions on Fe XIII.

Correspondence and requests for materials should be addressed to J.H.M.M.S. (e-mail: jschmitt@hs.uni-hamburg.de).

Evidence for ubiquitous strong electron–phonon coupling in high-temperature superconductors

A. Lanzara^{*†}, P. V. Bogdanov^{*}, X. J. Zhou^{*}, S. A. Kellar^{*}, D. L. Feng^{*}, E. D. Lu[†], T. Yoshida[‡], H. Eisaki^{*}, A. Fujimori[‡], K. Kishio[§], J.-I. Shimoyama[§], T. Noda^{||}, S. Uchida^{||}, Z. Hussain[†] & Z.-X. Shen^{*}

^{*} Department of Physics, Applied Physics and Stanford Synchrotron Radiation Laboratory, Stanford University, Stanford, California 94305, USA

[†] Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

[‡] Department of Physics; and [§] Department of Applied Chemistry, University of Tokyo, Bunkyo-ku, Tokyo 113-8656, Japan

^{||} Department of Superconductivity, University of Tokyo, Yayoi 2-11-16, Bunkyo-ku, Tokyo 133, Japan

Coupling between electrons and phonons (lattice vibrations) drives the formation of the electron pairs responsible for conventional superconductivity¹. The lack of direct evidence for electron–phonon coupling in the electron dynamics of the high-transition-temperature superconductors has driven an intensive search for an alternative mechanism. A coupling of an electron with a phonon would result in an abrupt change of its velocity and scattering rate near the phonon energy. Here we use angle-resolved photoemission spectroscopy to probe electron dynamics—velocity and scattering rate—for three different families of copper oxide superconductors. We see in all of these materials an abrupt change of electron velocity at 50–80 meV, which we cannot explain by any known process other than to invoke coupling with the phonons associated with the movement of the oxygen atoms. This suggests that electron–phonon coupling strongly influences the electron dynamics in the high-temperature superconductors, and must therefore be included in any microscopic theory of superconductivity.

We investigated the electronic quasiparticle dispersions in three different families of hole-doped copper oxides: $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$ (Bi2212) and Pb-doped Bi2212 (Pb-Bi2212), Pb-doped $\text{Bi}_2\text{Sr}_2\text{CuO}_6$ (Pb-Bi2201) and $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ (LSCO). Except for the Bi2201 (overdoped, transition temperature $T_c = 7 \text{ K}$) data and that in Fig. 3b, recorded at beam-line 5.4 of the Stanford Synchrotron Radiation Laboratory (SSRL), all the data were recorded at the Advanced Light Source (ALS), as detailed elsewhere².

The top panels of Fig. 1 report the dispersions—derived from the momentum distribution curves, MDCs—along the $(0, 0) - (\pi, \pi)$ direction for LSCO (Fig. 1a) and Bi2212 (Fig. 1b) in the superconducting state, and for Pb-Bi2201 in the normal state (Fig. 1c), versus the rescaled momentum, k' . We calculate k' by normalizing to one the momentum k relative to the Fermi momentum k_F , $(k - k_F)$, at the binding energy $E = 170 \text{ meV}$. A 'kink' in the dispersion around 50–80 meV, indicated by thick arrows in the figure, is the many-body effect of interest and has been previously reported^{2–5}. The direct comparison suggests a similar phenomenon in different systems (although details differ) and at different dopings, with the effect getting stronger in the underdoped region. These results put a strong constraint on the nature of the effect. The similar energy scale of the excitation (50–80 meV) in systems with very different gap energy, ranging from 10–20 meV for LSCO and Bi2201, to 30–50 meV for Bi2212, rules out the superconducting gap as the origin. The data also rule out the proposed explanation⁶ in terms of coupling with the magnetic mode at 41 meV (ref. 7), because of its temperature dependence and its ubiquity. The phenomenon is observed in LSCO where the magnetic mode does not exist. It is also observed well above the transition temperature T_c

in all cases, but the magnetic mode sets in at T_c for optimal and overdoped samples⁷ (for example, the 30 K data from overdoped Pb-Bi2201 are measured at six times T_c). These considerations, based on direct experimental evidence, leave phonons as the only possible candidates.

The phonon interpretation receives strong support from a direct comparison between photoemission results and neutron-scattering data on LSCO. As shown by the red arrow and the shaded area in Fig. 1a, the energy of the zone boundary in-plane oxygen-stretching longitudinal optical (LO) phonon, identified by neutrons as being strongly coupled to charge^{8,9}, coincides with the 'kink' energy in our data—barring a minor correction due to the small superconducting gap of 10 meV. We identify this mode as the highest energy phonon that contributes strongly to the ARPES (angle-resolved photoemission spectroscopy) data. However, other phonon contributions should be considered as well; for example, the apical mode¹⁰, which is related to the in-plane oxygen-stretching mode. Additional support for the phonon model comes also from an analysis of temperature dependence. In the lower panels of Fig. 1 we report the evolution of the 'kink' above and below the transition temperature for LSCO (Fig. 1d) and Bi2212 (Fig. 1e). In both cases, the 'kink' clearly persists above the transition temperature as expected for electron–phonon interaction, but a thermal broadening is present.

Figure 2a–c shows data from energy distribution curves (EDCs) of Bi2212 along the $(0, 0) - (\pi, \pi)$ direction at different dopings; also shown for comparison are corresponding data for the Be (0001) surface state^{11,12} (Fig. 2d)—in which the electron–phonon interaction is known to be strong—and simulated spectra (Fig. 2e). A

strong similarity among the raw spectra can be observed. In the case of Be, the 65-meV phonon sets a scale below which a sharp quasiparticle is possible. This leads to a 'dip' or 'break' in the spectra near the phonon energy, indicated by the vertical dashed line. This feature is well reproduced in the simple simulation, where a quasiparticle is isotropically coupled to a collective mode near 70 meV (refs 6, 13). The Bi2212 data show also a clear resemblance, with a $(\pi, 0)$ LO-phonon near 55 meV (ref. 9), but pushed up by the superconducting gap. All the data suggest that there is an energy scale close to the expected phonon frequency. Therefore, the notable similarity between the Bi2212 data and that of a known phonon case strongly supports the phonon explanation.

In Fig. 3a we report the quasiparticle width of Bi2212 for different doping, obtained using a typical procedure². As shown before^{2,3}, the width decreases more rapidly below the relevant energy, which corresponds to the kink position in the dispersion. This again is consistent with the case of an electron–phonon coupled system, complementing dispersion data in Fig. 1. The drop in the scattering rate near the relevant energy is also seen in optics experiments¹⁴. Figure 3b reports the 'dip' energy of ARPES spectra near $(\pi, 0)$ as a function of doping. The data are consistent with the idea that the 55-meV LO phonon energy sets the lower bound⁹. A realistic model—which considers the change of the $(\pi, 0)$ band position with doping, bilayer splitting (which severely distorts $(\pi, 0)$ spectra^{15,16}), electron–phonon interaction and energy gap—will be needed to understand the specifics of the data, including the small decrease of the dip energy with doping.

In a simple electron–phonon model, the ratio of the velocities of the bare and the dressed electrons is $(\lambda + 1)$, where λ measures

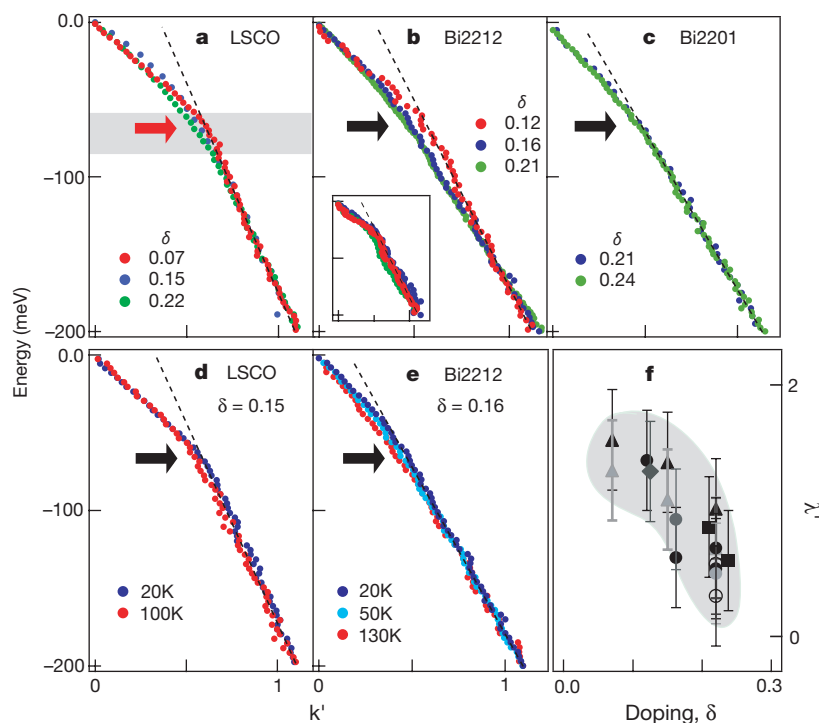


Figure 1 Ubiquity of a sudden change ('kink') in the dispersion. Top panels are plots of the dispersion (derived from the momentum distribution curves) along $(0, 0) - (\pi, \pi)$ (except panel **b** inset, which is off this line) versus the rescaled momentum K' for different samples and at different doping levels. **a–c**, Doping (δ) dependence of LSCO (at 20 K; **a**), Bi2212 (superconducting state, 20 K; **b**), and Bi2201 (normal state, 30 K; **c**). Dotted lines are guides to the eye. The kink position in **a** is compared with the phonon energy at $q = (\pi, 0)$ (thick red arrow) and the phonon width and dispersion (shaded area) from neutron data⁸. The doping was determined from the T_c versus doping universal curve. Inset in **b**, dispersions off the $(0, 0) - (\pi, \pi)$ direction, showing also a sharpening of the

kink on moving away from the nodal direction. The black arrows indicate the position of the kink in the dispersions. **d,e**, Temperature dependence of the dispersions for LSCO (**d**, optimally doped) and Bi2212 (**e**, optimally doped). **f**, Doping dependence of λ' (see text) along the $(0, 0) - (\pi, \pi)$ direction as a function of doping. Data are shown for LSCO (filled triangles) and NdLSCO (1/8 doping; filled diamonds), Bi2201 (filled squares) and Bi2212 (filled circles in the first Brillouin zone, and unfilled circles in the second zone). The different shadings represent data obtained in different experimental runs. Blue area is a guide to the eye.

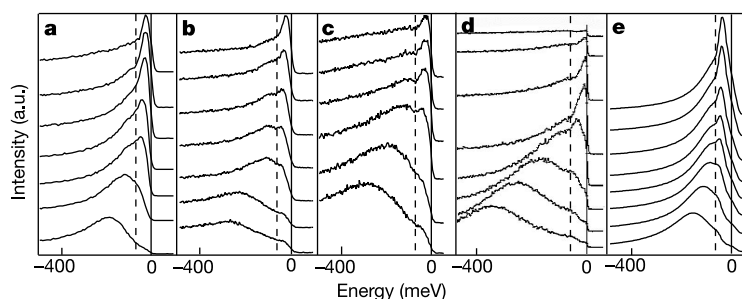


Figure 2 Double-peak features in the photoemission spectra due to strong electron–phonon interaction. **a–c**, Raw energy distribution curves (EDCs) along the $(0, 0) \rightarrow (\pi, \pi)$ direction for overdoped Pb-Bi2212 ($T_c = 70$ K; **a**), for optimally doped Bi2212 ($T_c = 91$ K; **b**) and for underdoped Bi2212 ($T_c = 84$ K; **c**). **d**, Raw EDCs for the Be(0001) surface¹¹.

e, EDCs of simulated spectra, obtained in the simple case of an isotropic coupling to a single phonon mode. The straight, vertical line highlights the zero-energy position, and a dashed line highlights the approximate ‘dip’ position, which is related to known phonon energy⁹ (see text).

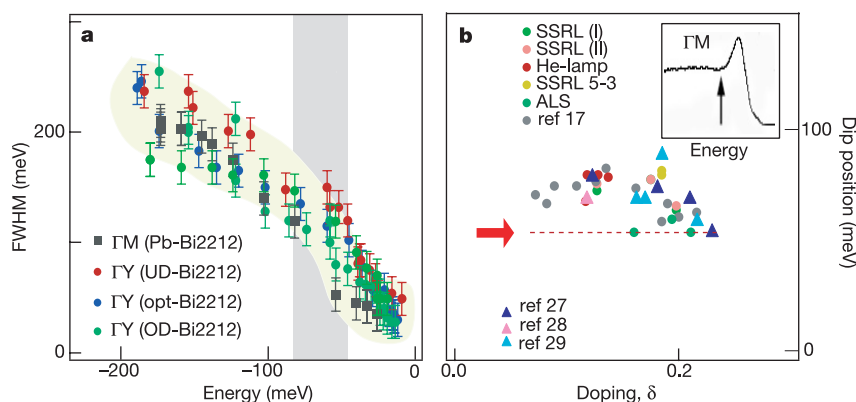


Figure 3 Comparison of the observed effect in photoemission, tunnelling and neutron data. **a**, Quasiparticle width along the ΓY direction $((0, 0) \rightarrow (\pi, \pi))$; circles) for Bi2212 and Pb-Bi2212, and along the ΓM direction $((0, 0) \rightarrow (\pi, 0))$; squares) for the superstructure-free Pb-Bi2212 system. **b**, Doping-dependence of the dip position for the Bi2212 samples, near the $(\pi, 0)$ region, as indicated in the inset of **b** by the arrow. The dip position has been obtained both from photoemission (circles) and tunnelling (triangles^{27–29}) data. The photoemission data were collected at SSRL, beamline 5.4 in

transmission mode (I) (green circles) and angular mode (II) (pink circles), and at SSRL, beamline 5.3 (yellow circles). The data at ALS were collected in angular mode (light green circles). Red circles represent the He-lamp data, collected with a Scienta analyser. Our data are compared with data points from ref. 17 (grey circles). The red arrow indicates the LO phonon energy at $(\pi, 0)$ ⁹, which sets the lower bound for the dip. FWHM, full-width at half-maximum; UD, underdoped; OD, overdoped; opt, optimum doping.

the electron–phonon coupling strength. We can extract a similar quantity from the ratio between the high-energy velocity above the phonon energy (which approximates the velocity without phonon interaction) and the dressed velocity below the phonon energy. The two velocities are determined by fitting the experimental dispersions with two straight lines. The coupling strength determined from the experimental data, λ' , is proportional to λ , but is an overestimate for two reasons. The first is that the high-energy dispersion overestimates the bare velocity. The second is that electron–electron scattering also influences the dispersion away from the phonon energy, especially in the high-energy part that we use as the bare velocity. In Fig. 1f we report the doping dependence of λ' along the (π, π) direction. An estimate of λ can alternatively be obtained by fitting the experimental data with a model (for example, the Debye model), or assuming the bare velocity as given from band calculations. In both cases the obtained values are smaller than λ' , but the decreasing trend with doping is identical.

In Fig. 4a, we show the velocity ratio for different materials as a function of the angle ϕ along the entire Fermi surface (Fig. 4b). Data obtained from different experimental runs are included in the figure (filled and unfilled symbols). In Fig. 4b we report the rescaled MDC-derived dispersions for Pb-Bi2212 taken with 55-eV photons along the two high-symmetry directions, ΓY $((0, 0) \rightarrow (\pi, \pi))$ and ΓM $((0, 0) \rightarrow (\pi, 0))$. The total change in the ratio with angle is slightly less than a factor of two. In particular, the data from both Pb-doped Bi2212 and Pb-doped Bi2201, where the superstructure effect near $(\pi, 0)$ is minimized, give confidence to the conclusion.

The conversion from velocity ratio to the coupling constant λ away from the (π, π) direction requires a correction due to the band structure effect (including bilayer-splitting) that increases away from the (π, π) direction^{15,16}, but the conclusion that the coupling is not very anisotropic remains true. No direct relation with the superconducting transition is observed, as can be seen in the same figure where data above (100 K) and below (20 K) the critical temperature are reported in the case of Pb-doped Bi2212. Both the angle dependence and the temperature dependence are consistent with the phonon interpretation, as phonons are more isotropic and will not disappear above T_c .

Although these experimental findings point toward phonons as the only possible model, some previous work needs to be addressed. The kink in the dispersion along the (π, π) direction and the dip in the EDCs near $(\pi, 0)$ has been attributed⁶ to the 41-meV magnetic mode (ref. 7). Comparing the doping dependence of the energy difference between the dip and the superconducting gap in photoemission data, with that of the magnetic mode in neutron data, a definitive proof for the magnetic mode as the origin of the ‘peak–dip–hump’ structure has been claimed¹⁷. On the other hand, the absence of an energy scale in the problem has been put forward⁴, and the kink in the dispersion has been overlooked⁴. The curving in the normal state dispersion has been attributed to the marginal Fermi liquid (MFL) self-energy¹⁸, and the presence of a sharp kink below T_c has been attributed to coupling to the (π, π) magnetic mode¹⁹.

Apart from its incompatibility with our data, the above magnetic

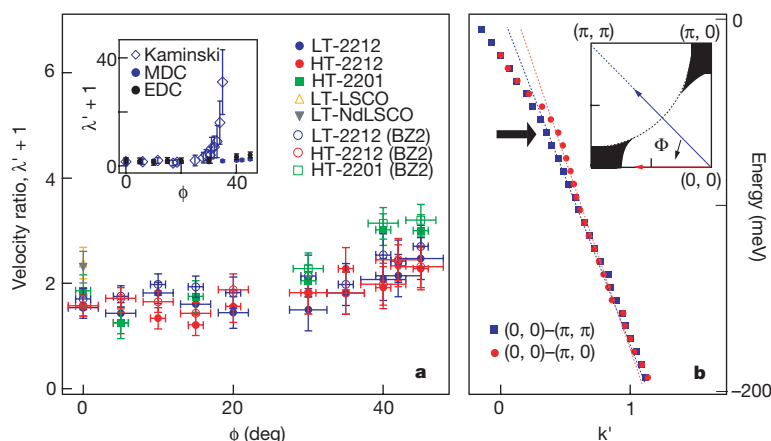


Figure 4 Momentum dependence of the velocity ratio above and below the critical temperature. The locations of the sections in the Brillouin zone are shown in the inset of **b**. ϕ is defined as the angle between the Γ -Y line and the line from Γ to the point on the Fermi surface considered. **a**, The velocity ratio versus angle ϕ along the Fermi surface. The open and filled symbols represent the values obtained fitting the MDC-derived dispersions in different experimental runs. HT and LT indicate the values of the velocity ratios obtained above and below the critical temperature. The HT data for Pb-Bi2212 were collected at 100 K (red circle) whereas the LT data were collected at 20 K (blue circle). The HT data for the Pb-Bi2201 system (squares) were collected at 30 K. The LT data (at 20 K)

mode model has serious weaknesses. First, the mode contains only a few per cent of the total spin fluctuation spectral weight²⁰, so it is an unlikely explanation for the very large change seen in EDC spectra above and below T_c . Second, the electronic model is difficult to implement self-consistently, as the mode will be altered by the interaction with electrons⁶. Third, the magnetic mode calculation claims to reproduce the experimental data that yield a huge change of the velocity ratio near the $(\pi, 0)$ region⁶. The inset of Fig. 4a compares the velocity ratio found in the present study with previous data³. The comparison shows good agreement for $\phi \leq 30^\circ$, but a large discrepancy is observed for $\phi \geq 30^\circ$. We attribute this difference to confusion caused by the superstructure, which complicates the data near the $(\pi, 0)$ region in the pure Bi2212 sample used in the previous study³. We have chosen superstructure-free Pb-Bi2212 for this investigation.

Our data and the above discussion rule out the magnetic mode as a possible model, and this also makes the 'MFL plus magnetic mode' interpretation an unsuitable explanation for the data. Furthermore, some of the normal-state dispersion data—for example, the 7% underdoped LSCO (Fig. 1a) measured at T_c where the kink is sharper—cannot be fitted within the MFL approach, even with an arbitrary choice of parameters. If the 7% LSCO data can only be explained by coupling to phonons, it is likely that a similar effect, observed at higher doping and temperatures, requires the same cause. Finally, this attempt to attribute the 'kink' in the dispersion, occurring at the same energy above and below T_c , to two different origins is not consistent with the data in Fig. 1d and e, where it can be seen that the effect washes out gradually with temperature.

The phonon model can explain all the aspects of the data attributed to the magnetic mode in a more natural way (N. Nagaosa *et al.*, personal communication). This approach has the added advantage of not suffering from the tiny spectral weight and the self-consistency issue present in the magnetic mode calculation. The phonon picture naturally explains the gradual temperature evolution observed in the MDC-derived dispersions.

Although electron–phonon interaction provides the only interpretation consistent with overall body of data, we need to mention a few caveats. The standard theory of electron–phonon coupling predicts a drop in resistivity at phonon frequency Ω , or a saturation

along the nodal direction for the LSCO and NdLSCO systems are also included (unfilled yellow and filled grey triangles). All velocity ratios are obtained for sections perpendicular to the Fermi surface at given ϕ . The inset of **a** compares the velocity ratios for Pb-Bi2212 as obtained from EDC (black filled circles) with data³ on pure Bi2212 (unfilled diamonds). In **b**, the rescaled MDC-derived dispersions, along the nodal and anti-nodal directions, are shown for the Pb-Bi2212. The small deviation above the phonon energy (indicated by the thick arrow) is caused by the difference in dispersion along the two directions, which are quite different if plotted with unscaled k .

above a temperature $T \approx (0.3 - 0.5)\Omega$. The drop at Ω has been observed and discussed in connection with the data in Fig. 3, although it was given an alternative interpretation²². The temperature dependence of the resistivity shows a complex doping dependence, with a saturation seen in the underdoped regime but not near the optimal and overdoped regimes²². An understanding of this complex behaviour requires going beyond the simple Fermi liquid theory, but is not incompatible with the presence of electron–phonon coupling²³. The lack of resistivity saturation is also seen in doped C_{60} compounds where the electron–phonon interaction is very strong²⁴. An improved theory needs to consider the strong electron–electron interaction, the pseudogap²⁵, and vertex correction, that suppress the large momentum transfer scatterings²⁶, making phonons hard to detect by resistivity experiments. Finally, we stress that electron–phonon interaction in strongly correlated materials is a largely unexplored topic, and we hope that our finding will stimulate more theoretical work. □

Received 13 February; accepted 5 June 2001.

1. Bardeen, J., Cooper, L. N. & Schrieffer, J. R. Theory of superconductivity. *Phys. Rev.* **108**, 1175–1204 (1957).
2. Bogdanov, P. V. *et al.* Evidence for an energy scale for quasiparticle dispersion in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8-\delta}$. *Phys. Rev. Lett.* **85**, 2581–2584 (2000).
3. Kaminski, A. *et al.* Renormalization of spectral lineshape and dispersion below T_c in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8-\delta}$. Preprint cond-mat/0004482 at (<http://xxx.lanl.gov>) (2000).
4. Valla, T. *et al.* Evidence for quantum critical behavior in the optimally doped cuprate $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8-\delta}$. *Science* **285**, 2110–2113 (1999).
5. Zhou, X. J. *et al.* Dual nature of the electronic structure of the stripe phase. *Phys. Rev. Lett.* **86**, 5578–5581 (2001).
6. Eschrig, M. & Norman, M. R. Neutron resonance: modeling photoemission and tunneling data in the superconducting state of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8-\delta}$. *Phys. Rev. Lett.* **85**, 3261–3264 (2000).
7. He, H. *et al.* Resonant spin excitation in an overdoped high temperature superconductor. Preprint cond-mat/0002013 at (<http://xxx.lanl.gov>) (2000).
8. McQueeney, R. J. *et al.* Anomalous dispersion of LO phonons in $\text{La}_{1.85}\text{Sr}_{0.15}\text{CuO}_4$ at low temperatures. *Phys. Rev. Lett.* **82**, 628–631 (1999).
9. Petrov, Y. *et al.* Phonon signature of charge inhomogeneity in high temperature superconductors $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$. Preprint cond-mat/0003414 at (<http://xxx.lanl.gov>) (2000).
10. Bianconi, A. *et al.* Determination of local lattice distortions in the CuO_2 plane of $\text{La}_{1.85}\text{Sr}_{0.15}\text{CuO}_4$. *Phys. Rev. Lett.* **76**, 3412–3415 (1996).
11. Hengsberger, M., Purdie, D., Segovia, P., Garnier, M. & Baer, Y. Photoemission study of a strongly coupled electron-phonon system. *Phys. Rev. Lett.* **83**, 592–595 (1999).
12. Lashell, S., Jensen, E. & Balasubramanian, T. Nonquasiparticle structure in the photoemission spectra from the $\text{Be}(0001)$ surface and determination of the electron self energy. *Phys. Rev. B* **61**, 2371–2374 (2000).

13. Scalapino, D. J. in *Superconductivity* (ed. Parks, R. D.) 449 (Marcel Dekker, New York, 1969)
14. Puchkov, A. V., Basov, D. N. & Timusk, T. The pseudogap state in high T_c superconductors: An infrared study. *J. Phys. Condens. Matter* **8**, 10049–10082 (1996).
15. Feng, D. L. *et al.* Bilayer splitting in the electronic structure of heavily overdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Phys. Rev. Lett.* **86**, 5550–5553 (2001).
16. Chuang, Y.-D. *et al.* Doubling of the bands in overdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ —probable evidence for c-axis bilayer coupling. Preprint cond-mat/0102386 at (<http://xxx.lanl.gov>) (2000).
17. Campuzano, J. C. *et al.* Electronic spectra and their relation to the (π, π) collective mode in high- T_c superconductors. *Phys. Rev. Lett.* **83**, 3709–3712 (1999).
18. Varma, C. M., Littlewood, P. B., Schmitt-Rink, S., Abrahams, E. & Ruckenstein, A. E. Phenomenology of the normal state of Cu-O high-temperature superconductors. *Phys. Rev. Lett.* **63**, 1996–1999 (1989).
19. Johnson, P. D. *et al.* On the doping and temperature dependence of the mass enhancement observed in the cuprate Bi2212 . Preprint cond-mat/0102260 at (<http://xxx.lanl.gov>) (2001).
20. Fong, H. F. *et al.* Polarized and unpolarized neutron scattering study of the dynamic spin susceptibility in $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Phys. Rev. B* **54**, 6708–6720 (1996).
21. Carbotte, J. P. *et al.* Coupling strength of charge carriers to spin fluctuations in high-temperature superconductors. *Nature* **401**, 354–356 (1996).
22. Takagi, H. *et al.* Systematic evolution of temperature-dependent resistivity in $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$. *Phys. Rev. Lett.* **69**, 2975–2978 (1992).
23. Allen, P. B. Anomalous versus conventional low energy properties of cuprate superconductors. *Comments Condens. Matter Phys.* **15**, 327–353 (1992).
24. Gunnarsson, O. Superconductivity in fullerenes. *Rev. Mod. Phys.* **69**, 575–606 (1997).
25. Batlogg, B. *et al.* Normal-state phase-diagram of $(\text{La, Sr})_2\text{CuO}_4$ from charge and spin dynamics. *Physica C* **235**, 130–133 (1994).
26. Kulić, M. L. Interplay of electron-phonon interaction and strong correlations: the possible way to high-temperature superconductivity. *Phys. Rep.* **338**, 1–264 (2000).
27. Renner, C. H. *et al.* Pseudogap precursor of the superconducting gap in under- and overdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$. *Phys. Rev. Lett.* **80**, 149–152 (1998); Observation of the low temperature pseudogap in the vortex core of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$. *Phys. Rev. Lett.* **80**, 3606–3609 (1998).
28. Cren, T. *et al.* Influence of disorder on the local density of states in high- T_c superconducting thin films. *Phys. Rev. Lett.* **84**, 147–150 (2000).
29. De Wilde, Y. *et al.* Unusual strong-effects in the tunneling spectroscopy of optimally doped and overdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$. *Phys. Rev. Lett.* **80**, 153–156 (1998).

Acknowledgements

We thank N. Nagaosa, D. J. Scalapino, R. Laughlin, D.-H. Lee, S. Kivelson, D. Bonn, K. A. Muller, P. Allen, N. P. Armitage, A. Damascelli and F. Ronning for discussions. The work at ALS and SSRL was supported by the Department of Energy's Office of Basic Energy Science, Division of Materials Science. The Stanford work was also supported by the NSF. A.L. thanks the Istituto Nazionale Fisica della Materia (INFM) and the University of Rome "La Sapienza" for support.

Correspondence and requests for materials should be addressed to: Z.-X.S. (e-mail: zxshen@stanford.edu).

Fragile-to-strong transition and polyamorphism in the energy landscape of liquid silica

Ivan Saika-Voivod*, Peter H. Poole* & Francesco Sciortino†

* Department of Applied Mathematics, University of Western Ontario, London, Ontario N6A 5B7, Canada

† Dipartimento di Fisica and Istituto Nazionale per la Fisica della Materia, Università di Roma La Sapienza, Piazzale Aldo Moro 2, I-00185, Rome, Italy

Liquid silica is the archetypal glass former, and compounds based on silica are ubiquitous as natural and man-made amorphous materials. Liquid silica is also the extreme case of a 'strong' liquid, in that the variation of viscosity with temperature closely follows the Arrhenius law as the liquid is cooled toward its glass transition temperature^{1,2}. In contrast, most liquids are to some degree 'fragile', showing significantly faster increases in their viscosity as the glass transition temperature is approached. Recent studies^{3–6,35,36} have demonstrated the controlling influence of the potential energy hypersurface (or 'energy landscape') of the liquid on the transport properties near the glass transition. But the origin of strong liquid behaviour in terms of the energy landscape has not yet been resolved. Here we study the static

and dynamic properties of liquid silica over a wide range of temperature and density using computer simulations. The results reveal a change in the energy landscape with decreasing temperature, which underlies a transition from a fragile liquid at high temperature to a strong liquid at low temperature. We also show that a specific heat anomaly is associated with this fragile-to-strong transition, and suggest that this anomaly is related to the polyamorphic behaviour of amorphous solid silica.

In a molecular dynamics computer simulation of an equilibrium liquid, the diffusion coefficient D is readily evaluated from the particle trajectories. Like the viscosity, D is a characteristic transport property whose deviations from the Arrhenius law serve to classify a liquid as strong or fragile. The theory of Adam and Gibbs (AG)⁷ states that D is related to the configurational entropy, S_c through

$$D = D_0 \exp(-A/TS_c) \quad (1)$$

where the parameters D_0 and A are commonly assumed to be independent of temperature, T . The entropy S_c quantifies the number of distinct configurational states explored by the liquid in equilibrium. It has been suggested that these states correspond to the 'basins' of the potential energy hypersurface (PES) sampled by the liquid^{8,9}. A basin is the set of points in phase space representing configurations having the same local minimum. The local minimum configuration is termed an inherent structure (IS), and is identified in simulation by a steepest descent minimization of the potential energy.

Following the thermodynamic formalism of Stillinger and Weber⁸, we can express the internal energy of the liquid as $E = e_{IS} + E_{\text{harm}} + E_{\text{anh}}$, where e_{IS} is the average inherent structure energy and the last two terms are the average contributions to E due to thermal excitations about the IS. The term E_{harm} is the average harmonic contribution determined from a quadratic approximation to E around each inherent structure minimum, and E_{anh} is the remaining, necessarily anharmonic contribution. The harmonic and anharmonic potentials characterize the shape of the basin.

If the shape of the basins does not depend on e_{IS} (a condition

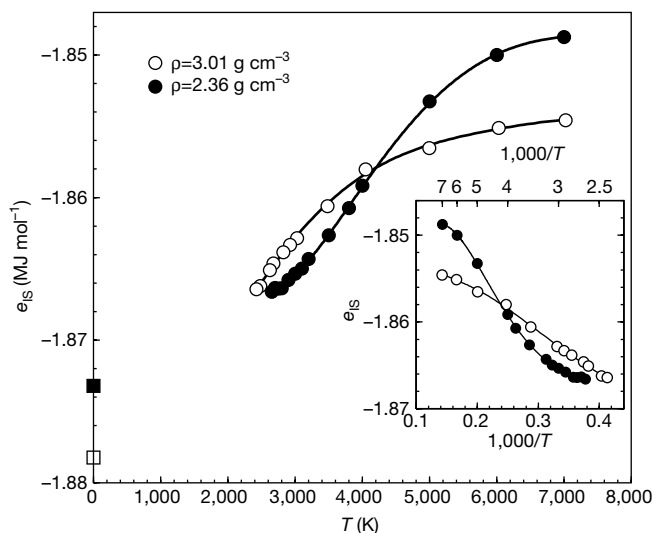


Figure 1 Variation of inherent structure energy e_{IS} with temperature T . Main panel: e_{IS} as a function of T along two isochores. At $T = 0$ we show the energy E_0 of the crystalline system at density $\rho = 2.36 \text{ g cm}^{-3}$ (filled square) and $\rho = 3.01 \text{ g cm}^{-3}$ (open square). E_0 is found by calculating the volume dependence of the potential energy at $T = 0$ of three crystal polymorphs of silica (stishovite, coesite and quartz), and then using the common tangent construction to determine the potential energy of the heterophase of coexisting crystals that would be the ground state at the required bulk value of ρ . Inset: e_{IS} for the same isochores as in the main panel, plotted as a function of $1/T$. Only the high-density data show a clear $1/T$ behaviour at low T .

satisfied at constant density ρ in the present study; see Methods), then S_c can be calculated along an isochore by integrating the T dependence of e_{IS} at constant volume V (ref. 3):

$$S_c = S_c^0 + \int_{T_0}^T \frac{1}{T'} \left(\frac{\partial e_{IS}}{\partial T'} \right)_V dT' \quad (2)$$

where S_c^0 is the value of S_c at a reference temperature $T = T_0$ (see Methods). Equation (2) highlights that the T dependence of S_c arises solely from changes in e_{IS} (refs 3, 8). Evaluation of S_c when the basin shape depends on e_{IS} is also feasible^{5,10,11}.

Some analyses of experimental data for silica suggest^{12,13} that the liquid may be fragile at very high T , and recent simulations^{14,15} of the BKS model of silica¹⁶ support this possibility. Simulations conducted between temperatures of 2,750 and 6,000 K have shown that at the onset of slow dynamics, as reflected for example by the presence of two-step relaxation in structural autocorrelation functions, BKS silica is a fragile liquid. At about 3,300 K BKS silica transforms to a strong liquid and the T dependence of all characteristic relaxation times follows the Arrhenius law¹⁵.

The relationship of such a 'fragile-to-strong' transition to the PES is not yet known. For a strong liquid that satisfies equation (1) the T dependence of S_c must approach a constant to recover Arrhenius behaviour. From equation (2), if S_c is constant, then so must be the variation of e_{IS} with T . This behaviour would be qualitatively different from that found in simulations of fragile liquids. For example, recent studies of a binary Lennard-Jones liquid have shown that e_{IS} is proportional to $-1/T$, a dependence consistent with a gaussian distribution of IS energies^{4,5}. When the distribution of inherent structure energy is gaussian, fragility has been shown to depend on the total number of IS states, the width of the gaussian, and the variation of the basin shape with e_{IS} (refs 5, 17). It is not known, however, if the PES of a liquid exhibiting a fragile-to-strong crossover is similarly characterized by a gaussian distribution of inherent structure energies but with parameters that differ from those of the fragile liquid, or if the PES is qualitatively different from that of a fragile liquid.

To address these questions, we conduct extensive equilibrium simulations of BKS silica over a large range of T and ρ to examine the fragile-to-strong transition, and to identify the energy landscape behaviour that underlies it. The results clarify our understanding both of the origins of silica's status as a strong liquid, and of the dynamical behaviour of a wider class of liquids that are to some degree silica-like, most notably water and silicon (S. Sastry, personal

communication), but also other systems such as BeF₂ (ref. 18). Indeed, the concept of a 'fragile-to-strong' transition was first proposed for the case of deeply supercooled water¹⁹.

Figure 1 shows the T dependence of e_{IS} along two isochores. The shape of the higher ρ isochore is similar to that found for fragile liquids. But at the lower ρ —which is close to the experimental ρ at ambient pressure— e_{IS} shows an inflection, passing from concave downwards at high T to concave upwards at low T . Figure 1 also shows the potential energy E_0 at $T = 0$ of the corresponding equilibrium crystalline system for the same ρ . As the energy of the lowest IS cannot be less than E_0 , E_0 sets a lower bound on e_{IS} . The value of E_0 relative to the measured e_{IS} curves confirms that an inflection occurs.

The inset of Fig. 1 shows that the relation $e_{IS} \propto -1/T$, the hallmark of a gaussian distribution of IS energies, is not obeyed along our lower ρ isochore. The breakdown of this relation does not arise from changes in the shapes of the basins sampled at different T . Indeed, we find that at constant ρ , the harmonic contribution to the vibrational entropy does not depend on the basin depth; our results are also consistent with the anharmonic contribution to the vibrational entropy being independent of basin depth (see Methods).

Like e_{IS} , S_c also has an inflection along our lower ρ isochore (Fig. 2). Our S_c data thus reveal the signature in the energy landscape of a fragile-to-strong transition. The rapid variation of landscape properties at high T corresponds to a fragile regime. As T decreases, the inflections of e_{IS} and S_c mark the onset of a regime in which the rate of change of these quantities decreases, consistent with the system's approach to the strong liquid limit.

In the range of T near the inflections of e_{IS} and S_c , we find (Fig. 3) that along both the high and low ρ isochores, D satisfies the AG relation within numerical error. That is, whether or not there is a change in the nature of the T dependence of S_c , the transport properties of the liquid adjust so as to maintain the AG relation, a finding that demonstrates the robustness of equation (1) (refs 5, 20, 21). As in simulations of binary Lennard-Jones liquids⁵ and water²¹, we also find that the slope of the isochores in Fig. 3 varies with ρ ; understanding this dependence is an important open question in the study of liquid dynamics using the PES.

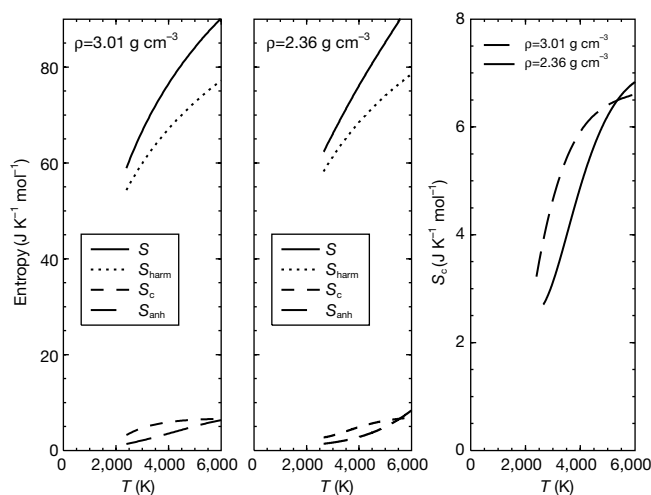


Figure 2 Entropy S and its component contributions as a function of temperature. Shown are isochores for $\rho = 3.01 \text{ g cm}^{-3}$ (left panel) and $\rho = 2.36 \text{ g cm}^{-3}$ (centre panel), and the configurational entropy S_c as a function of T along two isochores (right panel).

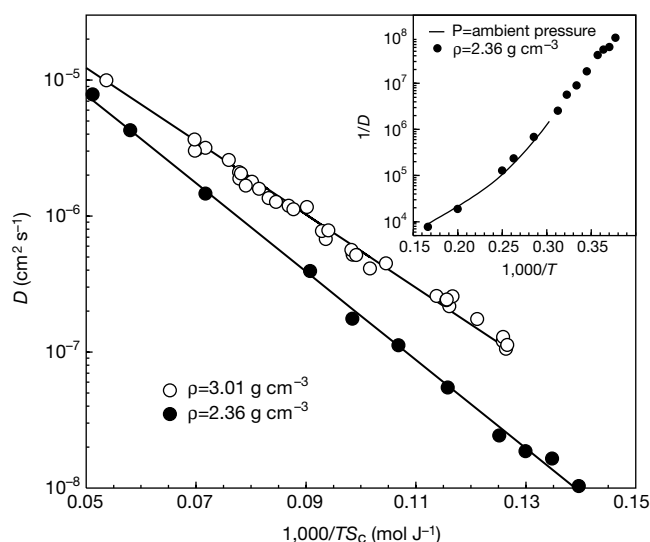


Figure 3 Relationship of diffusion coefficient D to temperature T and configurational entropy S_c . Main panel: $\log D$ versus $1/TS_c$ along two isochores. Inset: Arrhenius plot of $1/D$ along the ambient pressure isobar, found by interpolating our isochoric data; and along the $\rho = 2.36 \text{ g cm}^{-3}$ isochore. Note the crossover from non-Arrhenius to Arrhenius behaviour observed both at constant ρ and at constant pressure, P . All D values reported here are for Si atoms.

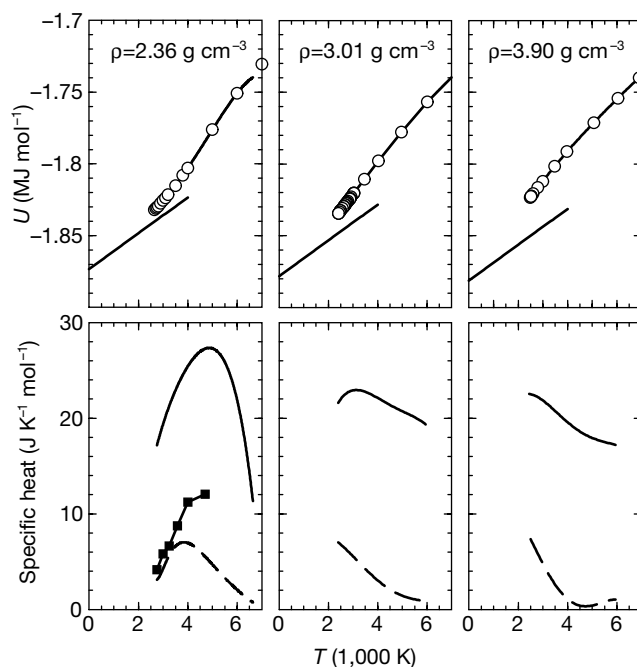


Figure 4 Potential energy U and isochoric specific heat C_V as a function of T . Upper panels: isochores of U . Lower panels: $C_V - (3/2)R$ (solid lines) and C_V^S (dashed lines) as a function of T . In the upper panels are shown lines of slope $\frac{3}{2}R$ whose values at $T = 0$ are the potential energies of the corresponding crystalline systems, calculated as described in

Fig. 1; these lines are lower bounds for U of the liquid. In the left-most lower panel are shown (filled squares) estimates of the configurational part of C_V recently evaluated by Scheidler *et al.*³⁴, calculated from the time dependence of the temperature fluctuations.

We note that for the present model, S_c is much smaller than the vibrational entropy ($S_{\text{harm}} + S_{\text{anh}}$) and of the same order as S_{anh} . Moreover, we find that for BKS silica the crystalline vibrational entropy differs significantly from the liquid vibrational entropy, as also found by Richet² in his extensive analysis of silicate melts. This confirms that for silica, the identification of the excess entropy (liquid entropy minus crystal entropy) with S_c may not be adequate^{22,23}. This highlights the value of finding S_c by the 'all-liquid route'²⁴ used here. We also note that estimates of the configurational entropy (expressed in the units used here) of a random tetrahedral network range from 0.83 to 1.94 J K⁻¹ mol⁻¹ (refs 25, 26). These estimates are consistent with our calculations, in that our S_c data are plausibly approaching this range from above at low T .

The influence of the energy landscape is sufficiently prominent to appear in the total thermodynamic properties²⁷. The isochoric specific heat C_V can be written as $C_V = C_V^{\text{IS}} + 3R + C_V^{\text{anh}}$, where each term is the derivative with respect to T of the corresponding contribution to E , and R is the gas constant. The inflection in the T dependence of e_{IS} corresponds to a maximum in C_V^{IS} (Fig. 4) that is the origin of a C_V anomaly, in the form of a peak, in the interval of T corresponding to the fragile-to-strong transition. An analogous C_V anomaly has recently been predicted to occur in the silica-like liquid BeF₂ (ref. 18), and in theoretical models designed to reproduce a fragile-to-strong transition²⁸. High- T experiments that test for this anomaly, although challenging, can thus directly seek the thermodynamic signature of the fragile-to-strong transition in these systems.

The peak that we find in C_V occurs at T near the temperature of maximum density of BKS silica, and near a maximum of the isothermal compressibility K_T predicted for this model²⁹. This interrelated set of liquid-state thermodynamic anomalies^{30,31} has been associated with the physics of polyamorphism in glasses—in other words, the rapid pressure-induced crossover of a low-density glass to a high-density glass³². Polyamorphism is observed experimentally during compression of silica glass, and the C_V maximum

found here may be the thermal anomaly corresponding to polyamorphism's mechanical anomaly. Along different isochores we find that the T at which the C_V maximum occurs decreases with increasing ρ , as would be expected for an anomaly related to polyamorphism in silica. The maxima of C_V and K_T demarcate a zone of a rapid, but continuous, crossover in the thermodynamic properties of the liquid. In BKS silica, there is evidence²⁹ that this crossover becomes at lower T a discontinuous liquid–liquid phase transition. Although it is conceivable that such a first-order phase transition develops in real silica under appropriate conditions of temperature and pressure, the phase transition may be pre-empted by the glass transition, rendering it unobservable as a liquid-state phenomenon.

Consideration of the combined behaviour that we find for dynamic and thermodynamic properties suggests that the fragile-to-strong transition is the dynamical transition corresponding to the thermodynamic crossover in the liquid that presages polyamorphism²⁸. More generally, our results provide a basis for considering all strong liquids as candidates for polyamorphism: a strong liquid arises through a fragile-to-strong transition, associated with which may be thermodynamic anomalies that are the liquid-state precursors to polyamorphism. □

Methods

Computer simulations

Results are based on molecular dynamics simulations of BKS silica. All data are obtained from systems of $N = 1,332$ atoms (888 O and 444 Si atoms), except for the $\rho = 2.36 \text{ g cm}^{-3}$ isochore, where we use 999 atoms to achieve equilibration more easily at low T . At $\rho = 2.36 \text{ g cm}^{-3}$, eight independent runs for each state point are made. All data reported here are for liquid states in equilibrium. Equilibration is confirmed by ensuring that runs are longer than the slowest relaxation time in the system as evaluated from the collective (coherent) density–density correlation function. The lowest- T runs exceed 80 ns. Simulations are carried out in the constant- (N, V, E) ensemble, and long-range electrostatic interactions are accounted for by Ewald summation. We evaluate e_{IS} by conducting conjugate gradient minimizations of up to 1,000 equilibrium liquid configurations and averaging the results. Note that our thermodynamic results are reported per mole of atoms, rather than per mole of SiO₂ molecules.

Evaluation of the total entropy

For a given ρ , S at $T = T_0 = 4,000$ K is calculated using thermodynamic integration. We first find the free energy difference between an ideal gas and a binary mixture Lennard-Jones (LJ) system at the chosen ρ by integrating the excess pressure along an isotherm from the high- V limit where the system behaves as an ideal gas. We then carry out a set of simulations at constant V and T that continuously convert the LJ system to BKS silica, by using a hybrid potential $\phi = \lambda\phi_{\text{BKS}} + (1 - \lambda)\phi_{\text{LJ}}$ (ref. 33). An appropriate thermodynamic integration from $\lambda = 0$ to 1 yields the free energy of BKS silica, from which S at T_0 and the chosen ρ is calculated. The value of S at other temperatures is found by further thermodynamic integration at constant V .

Evaluation of the vibrational entropy

We evaluate S_{vib} from the spectrum of eigenfrequencies ν (that is, the vibrational density of states) calculated from the inherent structures at each T , using $S_{\text{vib}} = k_B \sum_{\nu} [1 - \log(h\nu/kT)]$, where k and h are Boltzmann's and Planck's constants, respectively. To obtain S_{vib} we use E_{anh} . We evaluate $E_{\text{anh}} = E - E_{\text{harm}} - e_{\text{IS}}$ and then fit E_{anh} with a polynomial constrained to be zero and have zero slope at $T = 0$. Assuming that the shapes of the basins do not depend on e_{IS} , S_{vib} may be calculated by thermodynamic integration using the fitted form of E_{anh} from $T = 0$ to the desired T . In terms of the above quantities, the integration constant in equation (2) is thus $S_0^e = S(T_0) - S_{\text{vib}}(T_0) - S_{\text{harm}}(T_0)$.

Isochoric invariance of basin shape

Our assumption that the shape of the basins is independent of e_{IS} along an isochore is based on two observations. First we find that the vibrational density of states (the ν spectrum), is independent of e_{IS} along an isochore. Second, the anharmonic energy of inherent structure configurations heated from $T = 0$ to a chosen T follows the $E_{\text{anh}} = E - E_{\text{harm}} - e_{\text{IS}}$ curve calculated from equilibrium simulations. This is only possible if the anharmonic character of the basins is also to a large extent independent of e_{IS} .

Received 12 April; accepted 18 June 2001.

- Angell, C. A. Relaxation in liquids, polymers and plastic crystals—Strong/fragile patterns and problems. *J. Non-Cryst. Solids* **131–133**, 13–31 (1991).
- Richet, P. Viscosity and configurational entropy of silicate melts. *Geochim. Cosmochim. Acta* **48**, 471–483 (1984).
- Sciortino, F., Kob, W. & Tartaglia, P. Inherent structure entropy of supercooled liquids. *Phys. Rev. Lett.* **83**, 3214–3217 (1999).
- Buechner, S. & Heuer, A. The potential energy landscape of a model glass former: thermodynamics, anharmonicities, and finite size effects. *Phys. Rev. E* **60**, 6507–6518 (1999).
- Sastry, S. The relationship between fragility, configurational entropy and the potential energy landscape of glass-forming liquids. *Nature* **409**, 164–167 (2001).
- Martinez, L.-M. & Angell, C. A. A thermodynamic connection to the fragility of glass-forming liquids. *Nature* **410**, 663–667 (2001).
- Adam, G. & Gibbs, J. H. On the temperature dependence of cooperative relaxation properties in glass-forming liquids. *J. Chem. Phys.* **43**, 139–146 (1965).
- Stillinger, F. H. & Weber, T. A. Packing structures and transitions in liquids and solids. *Science* **225**, 983–989 (1984).
- Stillinger, F. H. A topographic view of supercooled liquids and glass formation. *Science* **267**, 1935–1939 (1995).
- Sciortino, F. & Tartaglia, P. Extension of the fluctuation-dissipation theorem to the physical aging of a model glass-forming liquid. *Phys. Rev. Lett.* **86**, 107–111 (2001).
- Starr, F. W., Sastry, S., La Nave, E., Scala, S., Stanley, H. E. & Sciortino, F. Thermodynamic and structural aspects of the potential energy surface of simulated water. *Phys. Rev. E* **63**, 041201 (2001).
- Rossler, E., Hess, K.-U. & Novikov, V. N. Universal representation of viscosity in glass forming liquids. *J. Non-Cryst. Solids* **223**, 207–222 (1998).
- Hess, K.-U., Dingwell, D. B. & Rossler, E. Parametrization of viscosity-temperature relations of aluminosilicate melts. *Chem. Geol.* **128**, 155–163 (1996).
- Barrat, J.-L., Badro, J. & Gillet, P. A strong to fragile transition in a model of liquid silica. *Molecular Simulation* **20**, 17–25 (1997).
- Horbach, J. & Kob, W. Static and dynamic properties of a viscous silica melt. *Phys. Rev. B* **60**, 3169–3181 (1999).
- Van Beest, B. W. H., Kramer, G. J. & van Santen, R. A. Force fields for silicas and aluminophosphates based on ab initio calculations. *Phys. Rev. Lett.* **64**, 1955–1958 (1990).
- Speedy, R. J. Relations between a liquid and its glasses. *J. Phys. Chem. B* **103**, 4060–4065 (1999).
- Hemmati, M., Moynihan, C. T. & Angell, C. A. Density maxima and minima, and water-like heat capacity and transport anomalies, in liquid BeF_2 . *J. Chem. Phys.* (in press).
- Angell, C. A. Water II is a strong liquid. *J. Phys. Chem.* **97**, 6339–6341 (1993).
- Speedy, R. J. The hard sphere glass transition. *Mol. Phys.* **95**, 169–178 (1998).
- Scala, A., Starr, F. W., La Nave, E., Sciortino, F. & Stanley, H. E. Configurational entropy and diffusivity of supercooled water. *Nature* **406**, 166–169 (2000).
- Johari, G. P. Contributions to the entropy of a glass and liquid, and the dielectric relaxation time. *J. Chem. Phys.* **112**, 7518–7523 (2000).
- Goldstein, M. Viscous liquids and the glass transition: sources of the excess heat capacity. *J. Chem. Phys.* **51**, 3728–3739 (1969).
- Coluzzi, B., Parisi, G. & Verrocchio, P. Thermodynamical liquid-glass transition in a Lennard-Jones binary mixture. *Phys. Rev. Lett.* **84**, 306–309 (2000).
- Bell, R. J. & Dean, P. The configurational entropy of vitreous silica in the random network theory. *Phys. Chem. Glasses* **9**, 125–127 (1968).
- Speedy, R. J. & Debenedetti, P. G. The distribution of tetravalent network glasses. *Mol. Phys.* **88**, 1293–1316 (1996).
- Stillinger, F. H., Debenedetti, P. G. & Sastry, S. Resolving vibrational and inherent structural contributions to isothermal compressibility. *J. Chem. Phys.* **109**, 3983–3988 (1998).

- Jagla, E. A. Fragile-strong transitions and polymorphism in glass forming fluids. *Mol. Phys.* **99**, 753–757 (2001).
- Saika-Voivod, I., Sciortino, F. & Poole, P. H. Computer simulation of liquid silica: equation of state and liquid-liquid phase transition. *Phys. Rev. E* **63**, 011202 (2001).
- Sastry, S., Debenedetti, P. G., Sciortino, F. & Stanley, H. E. Singularity-free interpretation of the thermodynamics of supercooled water. *Phys. Rev. E* **63**, 6144–6154 (1996).
- Rebelo, L. P. N., Debenedetti, P. G. & Sastry, S. Singularity-free interpretation of the thermodynamics of supercooled water. II. Thermal and volumetric behavior. *J. Chem. Phys.* **109**, 626–633 (1998).
- Poole, P. H., Grande, T., Angell, C. A. & McMillan, P. F. Polymorphic phase transitions in liquids and glasses. *Science* **275**, 322–323 (1997).
- Mezei, M. & Beveridge, D. L. Free energy simulations. *Ann. NY Acad. Sci.* **482**, 1–23 (1986).
- Scheidler, P., Kob, W., Latz, A., Horbach, J. & Binder, K. Frequency-dependent specific heat of silica. *Phys. Rev. B* **63**, 104204 (2001).
- Sastry, S., Debenedetti, P. G. & Stillinger, F. H. Signatures of distinct dynamical regimes in the energy landscape of a glass-forming liquid. *Nature* **393**, 554–557 (1998).
- Debenedetti, P. G. & Stillinger, F. H. Supercooled liquids and the glass transition. *Nature* **410**, 259–267 (2001).

Acknowledgements

We thank C. A. Angell, W. Kob, S. Sastry and R. Speedy for discussions. I.S.-V. and P.H.P. thank NSERC (Canada) for financial support, and SHARCNET for computing resources. F.S. acknowledges support from the INFN 'Iniziativa Calcolo Parallelo' and PRE-HOP and from MURST PRIN 2000.

Correspondence and requests for materials should be addressed to P.H.P. (e-mail: poole@cmrg.apmaths.uwo.ca).

Growth dynamics of pentacene thin films

Frank-J. Meyer zu Heringdorf, M. C. Reuter & R. M. Tromp

IBM T.J. Watson Research Center, Yorktown Heights, PO Box 218, New York 10598, USA

The recent demonstration of single-crystal organic optoelectronic devices has received widespread attention^{1–4}. But practical applications of such devices require the use of inexpensive organic films deposited on a wide variety of substrates. Unfortunately, the physical properties of these organic thin films do not compare favourably to those of single-crystal materials. Moreover, the basic physical principles governing organic thin-film growth and crystallization are not well understood. Here we report an *in situ* study of the evolution of pentacene thin films, utilizing the real-time imaging capabilities of photoelectron emission microscopy. By a combination of careful substrate preparation and surface energy control, we succeed in growing thin films with single-crystal grain sizes approaching 0.1 millimetre (a factor of 20–100 larger than previously achieved), which are large enough to fully contain a complete device. We find that organic thin-film growth closely mimics epitaxial growth of inorganic materials, and we expect that strategies and concepts developed for these inorganic systems will provide guidance for the further development and optimization of molecular thin-film devices.

'Plastic transistors' offer possibilities for flexible displays, and all-plastic smart cards and badges, as well as light-emitting diodes and lasers^{1,5–8}. Pentacene ($\text{C}_{14}\text{H}_{22}$), a chain-like aromatic molecule composed of five benzene rings, is among the most promising materials. Recent progress in organic electronics has focused on the exploration of new devices in single-crystal materials^{1–4,9}. However, electrical properties of polycrystalline films are inferior to those of single-crystal materials, and—because the mobility in single-crystal bulk material is higher⁵ than the values reported for organic thin-film transistors (OTFTs)—improvement of the film quality is mandatory¹⁰.

The different types of electrically active traps in pentacene devices have been analysed recently¹¹. Defects inside grains of polycrystalline films result in Fermi-level pinning, and primarily affect the transistor threshold voltage. Grain boundaries, in contrast, generate charge barriers that have to be overcome by thermionic emission. This does not only explain gate-voltage-dependent mobilities in OTFTs^{12,13}, but also implies that improved device performance may be expected if the grain-boundary density can be reduced by improvements in the thin-film growth process.

Previous studies of the growth of pentacene thin films include X-ray investigations of the molecular order⁶, atomic force microscopy⁷ and low-coverage scanning tunnelling microscopy¹⁴, as well as theoretical studies of the adsorption of single pentacene molecules on surfaces¹⁵. The results show that the molecule initially lies flat on the surface¹⁴. When several molecules have formed a stable island, they tilt out of the surface plane, and at room temperature form a unique thin-film phase⁶. The triclinic pentacene bulk structure¹⁶ is formed only during growth at elevated temperature².

Here we use low-energy electron microscopy (LEEM)^{17–19} to investigate the growth of pentacene thin films during *in situ* deposition. The IBM LEEM II instrument²⁰ was mostly operated in the 'photoelectron emission microscopy' mode (PEEM), using photoelectrons generated by ultraviolet radiation from a mercury discharge lamp. This allows a field of view of up to 65 μm with a resolution of 125 nm, limited by the pixel resolution of the video camera. The ultraviolet radiation causes the pentacene to desorb on a timescale of minutes. A computer-controlled shutter limits the ultraviolet exposure to 1 second when a PEEM image is taken once every minute, ensuring that film damage is negligible in our experiments.

Pentacene was deposited from a quartz crucible heated to $\sim 250^\circ\text{C}$. Clean material was obtained by carefully outgassing the pentacene in ultrahigh vacuum before deposition. Without this procedure, charge traps were introduced into the pentacene film as observed with LEEM. In a number of experiments, the sample was exposed to cyclohexene before the growth of pentacene films on it. Cyclohexene (C_6H_{10}), a liquid at room temperature, was purified by a repeated number of freeze–pump–thaw cycles. All depositions were done at room temperature.

In the PEEM images, contrast between pentacene in the first few molecular layers arises from differences in electronic structure. The first molecular layer appears bright on a dark Si surface (Fig. 1a). In Fig. 1b, the second molecular layer appears dark on the first layer, and in Fig. 1c the third layer appears darker still. With increasing film thickness, contrast disappears as the electronic properties of the surface layer approach those of the bulk.

During the initial stages of growth, stable two-dimensional islands nucleate on the surface. The nucleation density depends on the deposition rate and on the preparation of the substrate. Whereas the nucleation density on clean silicon surfaces is of the order of $10^{-3} \mu\text{m}^{-2}$, it can easily be 100 times larger on SiO_2 . Figure 2 shows the evolution of three pentacene islands on $\text{Si}(001)$ during growth at a rate of 10^{-2} monolayers per minute (one monolayer

(ML) is defined as a single molecular layer of crystalline pentacene with a thickness of $\sim 15 \text{ \AA}$). Low-energy electron microdiffraction shows that each island forms a pentacene single crystal, and that the three crystals in Fig. 2a are rotated by a random angle with respect to each other. With increasing coverage the islands grow (Fig. 2b) and a second layer nucleates (arrows in Fig. 2c), while the first-layer islands still avoid touching. In Fig. 2d, trenches between islands are almost filled in, leaving grain boundaries that will eventually affect the electrical transport properties.

The fractional layer coverage during growth on clean $\text{Si}(001)$ is plotted in Fig. 3a. After deposition is started, it takes several minutes before nucleation occurs. During this dead-time, about one-fifth of a monolayer of pentacene is deposited. Comparison with results from scanning tunnelling microscopy¹⁰ indicates that pentacene molecules are adsorbed flat on the surface during the dead-time, and that the islands shown in Fig. 2a nucleate on top of this layer. We believe that this initial layer reacts with, and becomes immobilized by, the high density of reactive dangling bonds on the clean $\text{Si}(001)(2 \times 1)$ surface. After nucleation, the first layer grows linearly with time until, at a coverage of 60%, islands nucleate in the second layer. Again, when the coverage in this layer reaches 60%, nucleation in the third layer takes place, showing steady-state layer-by-layer growth.

Removal of the clean-surface dangling bonds should remove the initial dead-time, and force immediate nucleation of the first layer. Recent experiments show that adsorption of cyclopentene on $\text{Si}(001)$ passivates all dangling bonds²¹, and that adsorption of cyclohexene on germanium gives similar results²². The layer coverage for growth of pentacene on cyclohexene-saturated $\text{Si}(001)$ is plotted in Fig. 3b. As expected, no dead-time is observed. Other substrates without dangling bonds, such as thermal SiO_2 grown *in situ*, also show no dead-time. In all three cases discussed above, the well defined and highly uniform initial surfaces result in much lower nucleation densities (that is, larger grain sizes) than previously observed. This improvement in grain size is in large measure due to the absence of heterogeneous nucleation in our experiments. In contrast, we find that growth on poorly controlled oxide surfaces is dominated by dense heterogeneous nucleation.

Growth of the initial molecular layers is controlled by a number of factors. The presence of reactive sites on the surface may give rise to an initial dead-time. Beyond that, the homogeneous nucleation density of the first crystalline molecular layer, N_1 , is determined by the ratio F/D_s , where F is the growth rate and D_s is the diffusion constant of pentacene on the substrate. Nucleation of the second layer takes place when the first layer islands exceed a critical size R_c —which depends on F , the diffusion constant of pentacene on pentacene D_p , and the excess diffusion barrier at the island step-edge, E_B (the so-called Ehrlich-Schwoebel barrier)²³. When $\pi R_c^2 N_1 < 1$, second-layer nucleation occurs before the first layer closes²⁴. As D_s and D_p are not identical, N_1 depends on the substrate, whereas R_c does not. As the initial nucleation density increases, the coverage at which a new layer nucleates changes from 0.6 for clean $\text{Si}(001)$ to 0.8 for SiO_2 . However, the resulting improvement in

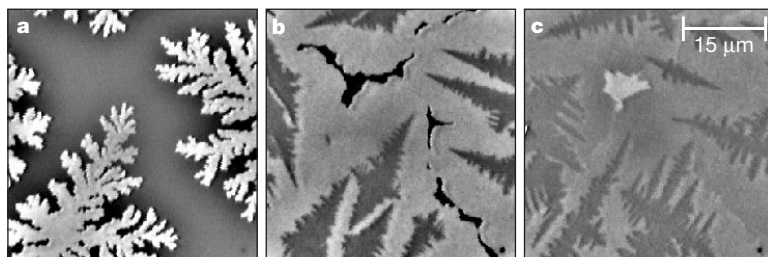


Figure 1 a–c, Development of the pentacene layer-by-layer contrast during deposition. **a**, One layer; **b**, two layers; **c**, three layers. Scale bar applies to all panels.

surface morphology with smoother films is offset by an increase in grain-boundary density. These two effects will have opposite influences on OTFT mobilities.

Despite the differences in the critical coverage for second-layer nucleation, the first layer of pentacene grows similarly on differently prepared samples. With sufficiently low nucleation density, a fractal island shape can always be observed. Fractal growth will occur when the molecular hopping time on the surface, τ_s , is much smaller than the hopping time at a step-edge, τ_e . From the finite width of the fractal arms in Fig. 1, it is clear²⁵ that transient diffusion at the step edges occurs immediately after the molecule sticks to the edge, and continues until it finds a suitable site with high coordination. But the complete lack of change in fractal shape over a period of 24 hours when growth is interrupted in ultrahigh vacuum indicates that pentacene molecules are indeed completely immobile, once adsorbed into the molecular layer.

An appropriate measure for the shape of islands in the first molecular layer is the fractal dimension, as determined by box counting^{26,27}. In Fig. 4a, the fractal dimension of the first pentacene layer is plotted versus coverage for several experiments, including usage of various substrates, application of cyclohexene, and different growth rates. Initially (when coverage $\Theta < 5 \times 10^{-2}$ ML), it is impossible to estimate the fractal dimension of the extremely small islands correctly, and box counting results in dimensions between zero and one. With increasing island size, however, the fractal dimension approaches a value of $D = 1.6$ – 1.7 and remains constant between 0.05 ML and 0.2 ML. At higher coverage coarsening effects dominate, the islands coalesce, and the fractal dimension converges to $D = 2$, as expected for a closed layer. The observed value of $D = 1.6$ – 1.7 agrees well with the expected dimension for classical diffusion-limited aggregation (DLA)²⁸.

To understand the dependence of the fractal dimension on coverage, we have performed a simulation in which molecules are deposited on a square lattice with 512×512 adsorption sites. Molecules can diffuse freely until they hit the edge of an island and stick. Nucleation of islands in the second layer is not included: all molecules deposited on top of the first layer diffuse to the boundary of the island, which acts as a perfect sink. For simplicity, we use periodic boundary conditions of the underlying lattice, and the initial number of islands is given as a starting parameter. After each simulation step the fractal dimension is calculated, and the

simulation is continued until the first layer is closed. The result is plotted in Fig. 4b.

The qualitative behaviour agrees well with that shown in Fig. 4a. Both curves show a plateau in the fractal dimension, and a continuous increase to $D = 2$ for the closed layer. While the result $0 < D < 1.5$ for low coverages is an artefact of the box counting, the plateau describes a DLA regime where dendrites grow without interaction. The somewhat lower plateau value of $D = 1.5$ – 1.6 for the simulation is a known result for DLA on a square lattice²⁹. When the diffusion fields of the islands begin to overlap, the shape of the dendrites changes, the edges smoothen, and the fractal dimension approaches $D = 2$.

Our results elucidate several important aspects of the growth of pentacene thin films. The first is nucleation and grain size. Previous experiments in the context of OTFT fabrication have shown high nucleation densities, with average grain sizes of about $1 \mu\text{m}$ —and, as noted earlier, electrically active traps at grain boundaries are detrimental to device performance. We have shown that the nucleation density strongly depends on the substrate. Grain sizes of pentacene grown at room temperature on clean Si(001) with or without cyclohexene, approach 0.1 mm ; that is, 100 times larger than in previous work. Using such large grains, an OTFT could be entirely contained inside a single grain; we expect single-crystal transport properties to be approached in such conditions. Work is under way to apply these findings in a well-controlled OTFT growth facility.

Second, our results show that the layer-by-layer growth phenomena observed here for pentacene can be understood in the general

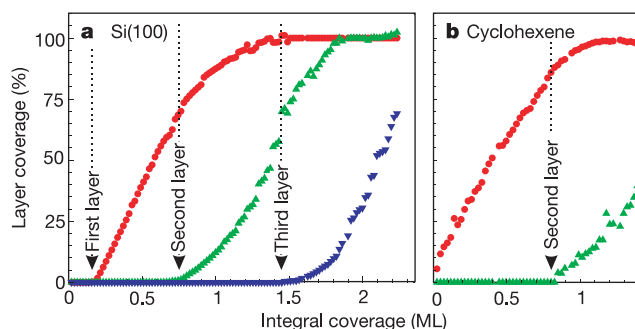


Figure 3 Layer coverage of pentacene during deposition on different substrates. **a**, Si(001); **b**, cyclohexene-saturated Si(001). The coverages in the first, second and third layer are plotted in red, green and blue, respectively.

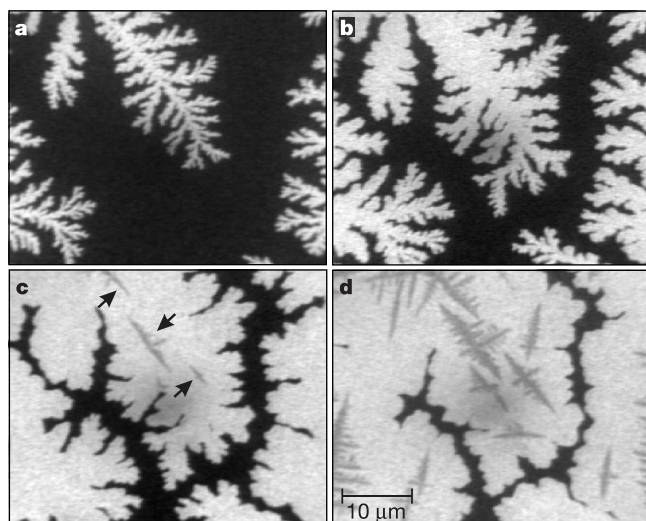


Figure 2 Evolution of the dendritic shape of pentacene islands on cyclohexene-saturated Si(001). **a**, $\Theta = 0.25 \text{ ML}$; **b**, $\Theta = 0.5 \text{ ML}$; **c**, $\Theta = 0.75 \text{ ML}$; **d**, $\Theta = 1 \text{ ML}$. Scale bar applies to all panels.

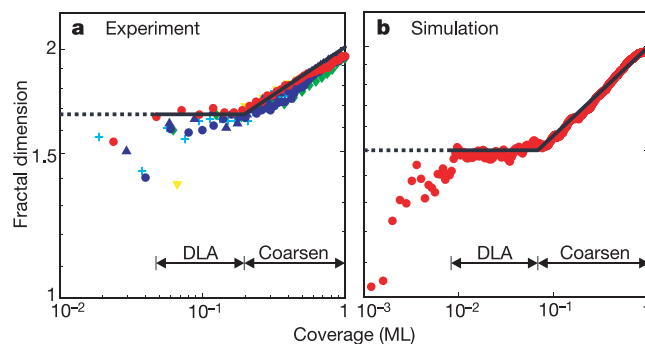


Figure 4 Coverage-dependent fractal dimension of single-molecular-layer pentacene islands on cyclohexene-saturated Si(001). **a**, Experiment; **b**, modified DLA (diffusion-limited aggregation) simulation. Independently of the experimental conditions, including growth rate variations and substrate surface preparation as illustrated by the different colours and symbols, the fractal dimension follows a universal curve.

framework developed in great detail for inorganic homo- and heteroepitaxial growth²³. Strategies developed to improve growth morphologies in metals³⁰ could be applied equally to these organic materials. Last, we have shown that the universal dynamic evolution of the first-layer fractal dimension is well explained by a growth mechanism of the DLA type, modified to take account of overlapping diffusion fields at higher coverage. We believe that the basic understanding of the growth of pentacene thin films afforded by these observations will contribute significantly to the development of organic thin-film devices and technology. □

Received 3 March; accepted 16 May 2001.

- Schön, J., Kloc, C., Dodabalapur, A. & Batlogg, B. An organic solid state injection laser. *Science* **289**, 599–601 (2000).
- Schön, J., Kloc, C. & Batlogg, B. Superconductivity in molecular crystals induced by charge injection. *Nature* **406**, 702–704 (2000).
- Schön, J., Kloc, C., Laudise, R. & Batlogg, B. Electrical properties of single crystals of rigid rodlike conjugated molecules. *Phys. Rev. B* **58**, 12952–12957 (1998).
- Schön, J., Dodabalapur, A., Kloc, C. & Batlogg, B. A light-emitting field-effect transistor. *Science* **290**, 963–965 (2000).
- Dimitrakopoulos, C., Purushothaman, S., Kyymissis, J., Callegari, A. & Shaw, J. Low-voltage organic transistors on plastic comprising high-dielectric constant gate insulators. *Science* **283**, 822–824 (1999).
- Dimitrakopoulos, C., Brown, A. & Pomp, A. Molecular beam deposited thin films of pentacene for organic field effect transistor applications. *J. Appl. Phys.* **80**, 2501–2508 (1996).
- Klauk, H. & Jackson, T. Pentacene organic thin-film transistors and ICs. *Solid State Technol.* **43**, 63–77 (2000).
- Klauk, H., Gundlach, D., Bonse, M. & Jackson, T. A reduced complexity process for organic thin film transistors. *Appl. Phys. Lett.* **76**, 1692–1694 (2000).
- Schön, J., Kloc, C., Bucher, E. & Batlogg, B. Efficient organic photovoltaic diodes based on doped pentacene. *Nature* **403**, 408–410 (2000).
- Gundlach, D. J., Lin, Y. Y., Jackson, T. N., Nelson, S. F. & Schlom, D. G. Pentacene organic thin-film transistors—molecular ordering and mobility. *IEEE Electron Device Lett.* **18**, 87–89 (1997).
- Schön, J. H. & Batlogg, B. Trapping in organic field-effect transistors. *J. Appl. Phys.* **89**, 336–342 (2001).
- Horowitz, G., Hajilaoui, R., Fichou, D. & El Kassmi, A. Gate voltage dependent mobility of oligothiophene field effect transistors. *J. Appl. Phys.* **85**, 3202–3206 (1999).
- Dimitrakopoulos, C. *et al.* Low-voltage, high-mobility pentacene transistors with solution-processed high dielectric constant insulators. *Adv. Mater.* **11**, 1372–1375 (1999).
- Kasaya, M., Tabata, H. & Kawai, T. Scanning tunneling microscopy and molecular orbital calculation of organic molecules adsorbed on the Si(100) 2x1 surface. *Surf. Sci.* **400**, 367–374 (1998).
- Yamaguchi, T. Electronic states of adsorbed organic molecule: pentacene on Si(001) 2x1 surface. *J. Phys. Soc. Jpn* **68**, 1321–1330 (1998).
- Campbell, R., Robertson, J. & Trotter, J. The crystal and molecular structure of pentacene. *Acta Crystallogr.* **14**, 705–711 (1961).
- Bauer, E. The resolution of the low energy electron reflection microscope. *Ultramicroscopy* **17**, 51–56 (1985).
- Teliops, W. & Bauer, E. An analytical reflection and emission UHV surface electron microscope. *Ultramicroscopy* **17**, 57–66 (1985).
- Bauer, E. Low energy electron microscopy. *Rep. Prog. Phys.* **57**, 895–938 (1994).
- Tromp, R., Mankos, M., Reuter, M., Ellis, A. & Copel, M. A new low energy electron microscope. *Surf. Rev. Lett.* **5**, 1189–1197 (1998).
- Hamers, R. *et al.* Cycloaddition chemistry of organic molecules with semiconductor surfaces. *Acc. Chem. Res.* **33**, 617–624 (2000).
- Lee, S., Hovis, J., Coulter, S., Hamers, R. & Greenleaf, C. Cycloaddition chemistry on germanium(001) surfaces: The adsorption and reaction of cyclopentene and cyclohexene. *Surf. Sci.* **462**, 6–18 (2000).
- Zhang, Z. & Lagally, M. Atomistic processes in the early stages of thin-film growth. *Science* **276**, 377–383 (1997).
- Terstoff, J., Denier van der Gon, A. & Tromp, R. Critical island size for layer by layer growth. *Phys. Rev. Lett.* **72**, 266–269 (1994).
- Zhang, Z., Chen, X. & Lagally, M. Bonding-geometry dependence of fractal growth on metal surfaces. *Phys. Rev. Lett.* **73**, 1829–1832 (1994).
- Feder, J. *Fractals* (Plenum, New York, 1988).
- Pickover, C. *Computers, Pattern, Chaos and Beauty—Graphics from an Unseen World* (St Martin's, New York, 1990).
- Witten, T. & Sander, L. Diffusion limited aggregation, a kinetic critical phenomenon. *Phys. Rev. Lett.* **47**, 1400–1403 (1981).
- Ball, R., Brady, R., Rossi, G. & Thompson, B. Anisotropy and cluster growth by diffusion-limited aggregation. *Phys. Rev. Lett.* **55**, 1406–1409 (1985).
- Wulfschkel, W. *et al.* Conventional and manipulated growth of Cu/Cu(111). *Surf. Sci.* **348**, 227–242 (1996).

Acknowledgements

We thank J. Shaw and C. Dimitrakopoulos for discussions. F.M.z.H. was supported by the Alexander von Humboldt Foundation under the Feodor-Lynen programme.

Correspondence and requests for materials should be addressed to F.M.z.H. (e-mail: meyerzh@us.ibm.com).

Ionic conductivity in crystalline polymer electrolytes

Zlatka Gadjeurova, Yuri G. Andreev, David P. Tunstall* & Peter G. Bruce

School of Chemistry, University of St Andrews, North Haugh, St Andrews, Fife KY16 9ST, UK

* School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews, Fife KY16 9SS, UK

Polymer electrolytes are the subject of intensive study, in part because of their potential use as the electrolyte in all-solid-state rechargeable lithium batteries¹. These materials are formed by dissolving a salt (for example LiI) in a solid host polymer such as poly(ethylene oxide) (refs 2–6), and may be prepared as both crystalline and amorphous phases. Conductivity in polymer electrolytes has long been viewed as confined to the amorphous phase above the glass transition temperature, T_g , where polymer chain motion creates a dynamic, disordered environment that plays a critical role in facilitating ion transport^{2,3,7–9}. Here we show that, in contrast to this prevailing view, ionic conductivity in the static, ordered environment of the crystalline phase can be greater than that in the equivalent amorphous material above T_g . Moreover, we demonstrate that ion transport in crystalline polymer electrolytes can be dominated by the cations, whereas both ions are generally mobile in the amorphous phase¹⁰. Restriction of mobility to the lithium cation is advantageous for battery applications. The realization that order can promote ion transport in polymers is interesting in the context of electronically conducting polymers, where crystallinity favours electron transport^{11,12}.

Polymer electrolytes form crystalline compounds at certain discrete compositions. For example, the crystalline compound $\text{PEO}_3\text{:LiCF}_3\text{SO}_3$ is obtained by dissolving LiCF_3SO_3 in poly(ethylene oxide) (PEO) with a composition corresponding to three ether oxygens per lithium ion. Polymer electrolytes may also be obtained as amorphous materials over a wide range of compositions and temperatures including, in some cases, the compositions at which crystalline complexes exist. Ionic conductivity occurs in the amorphous phase above T_g , where ion transport is induced by local motion of polymer chain segments repeatedly creating new coordination sites into which the ions may then migrate. For example, NMR studies of ionic motion within the $\text{PEO}\text{:LiCF}_3\text{SO}_3$ system, at compositions where the crystalline 3:1 complex and the amorphous phase coexist above T_g , demonstrate that ion transport occurs in the amorphous phase⁷.

Crystalline polymer electrolytes have received relatively little attention, mainly because they were not considered to support ionic conductivity. We have carried out an extensive study of the structural chemistry of crystalline polymer electrolytes, developing a method by which crystal structures of molecular solids may be determined *ab initio* from powder diffraction data^{13–15}. We have applied this and other methodologies to determine the structures of several crystalline polymer electrolytes^{14–17}. Of particular relevance are the structure determinations of the crystalline complexes formed with six ether oxygens per cation: $\text{PEO}_6\text{:LiPF}_6$, $\text{PEO}_6\text{:LiAsF}_6$ and $\text{PEO}_6\text{:LiSbF}_6$ (ref. 18). All three compounds are isostructural; the structure of $\text{PEO}_6\text{:LiAsF}_6$ is shown in Fig. 1 (ref. 15). Pairs of PEO chains fold to form cylindrical tunnels within which the Li^+ ions are located and coordinated by the ether oxygens. The anions are located outside these tunnels in the interchain space and do not coordinate the cations. The structure suggests that Li^+ ion transport along the tunnels may be possible in the crystalline 6:1 complexes.

To compare ionic conductivity in the static, ordered environment

of a crystalline polymer electrolyte with the dynamic and disordered environment of an amorphous polymer electrolyte above T_g , we prepared crystalline and amorphous forms of $\text{PEO}_6\text{:LiSbF}_6$. By combining LiSbF_6 with methoxy end-capped PEO of low weight-averaged molecular mass ($M_w = 1,000$) a crystalline 6:1 phase was obtained free from contamination by any amorphous material that would coexist with the crystalline phase at high M_w (above the entanglement limit for PEO of 3,200). Powder X-ray diffraction showed that the crystal structures obtained using low (1,000) and high (100,000) M_w values are the same^{15,18}. The peak widths indicate that the low- M_w material is highly crystalline, with the structure being coherent on a length scale in excess of 1,000 Å. By combining LiSbF_6 with high- M_w PEO (100,000) a single 6:1 amorphous phase was obtained (see Methods).

Variation of conductivity as a function of temperature for the crystalline and amorphous $\text{PEO}_6\text{:LiSbF}_6$ materials is shown in Fig. 2. The temperature range was selected to ensure that the conductivity data for amorphous $\text{PEO}_6\text{:LiSbF}_6$ corresponded to the material above T_g (NMR) and that the crystalline and amorphous phases were stable. The melting point for $\text{PEO}_6\text{:LiAsF}_6$ ($M_w = 100,000$) is 130 °C and drops by some 30 °C on reducing the weight-averaged molecular mass to 1,000. The low- M_w PF_6 and SbF_6 complexes both decompose at ~70 °C on extended heating, before they melt. The glass transition temperature was determined by NMR spectroscopy¹⁹. The ^1H resonance associated with the PEO chains was monitored as a function of temperature (Fig. 3). The rapidly narrowing linewidth is directly associated with the onset of PEO chain motion which serves to average the dipolar interactions and hence narrow the line. These results indicate a T_g for $\text{PEO}_6\text{:LiSbF}_6$ of around 240 K. It is evident from Fig. 2 that ionic conductivity in crystalline $\text{PEO}_6\text{:LiSbF}_6$ is higher than for the same composition in the amorphous state, even above T_g . The crystalline phase reaches a conductivity more than one order of magnitude higher than the amorphous phase at the lowest temperatures. The use of different molecular masses for crystalline ($M_w = 1,000$) and amorphous ($M_w = 100,000$) $\text{PEO}_6\text{:LiSbF}_6$ does not compromise the comparison of conductivity because, as pointed above, the crystal structure is the same regardless of M_w . We have therefore achieved a direct

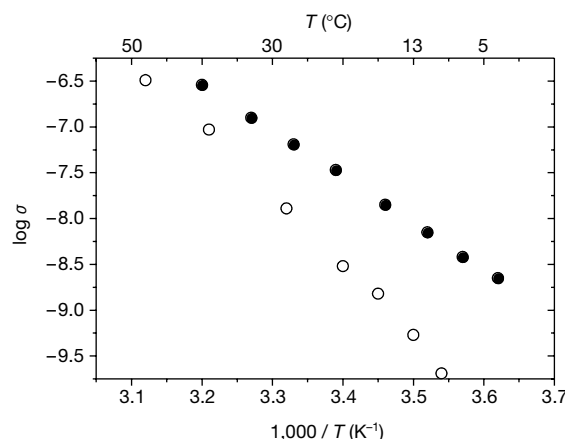


Figure 2 Ionic conductivity σ (S cm^{-1}) of amorphous (open circles) and crystalline (filled circles) $\text{PEO}_6\text{:LiSbF}_6$ as a function of temperature.

comparison of ionic conductivity in a static crystalline PEO environment with that in an amorphous PEO environment above T_g and for the same 6:1 composition. Comparison of amorphous and crystalline phases below the entanglement limit is not possible because an amorphous material with $M_w = 1,000$ would, above T_g , exhibit conductivity like that of a conventional liquid electrolyte in which the polymer solvent molecules migrate with the ions. Our aim is to compare conductivity in the crystalline phase with the unique mechanism of polymer electrolyte conductivity that exists above the entanglement limit. Although we have chosen $M_w = 100,000$ for the amorphous phase, it has been demonstrated²⁰ that the conductivity barely increases on reducing M_w towards the entanglement limit of 3,200.

It is important to demonstrate the lack of amorphous phase in the crystalline material and *vice versa*. We collected ^7Li NMR data at room temperature for the crystalline and amorphous forms of $\text{PEO}_6\text{:LiSbF}_6$. In each case a single line was observed, but the

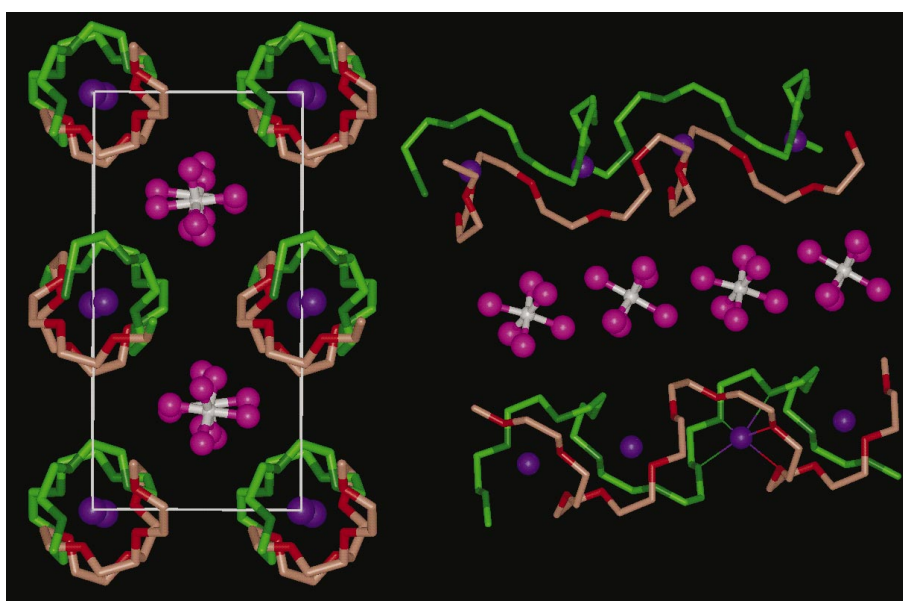


Figure 1 The structure of the polymer electrolyte $\text{PEO}_6\text{:LiAsF}_6$. Left, view of the structure along the chain axis showing rows of Li^+ ions perpendicular to the page. Right, view of the structure showing the relative position of the chains and their conformation (hydrogens not shown). Blue spheres, lithium; white spheres, arsenic; pink spheres,

fluorine; light green, carbon in chain 1; dark green, oxygen in chain 1; light red, carbon in chain 2; dark red, oxygen in chain 2. Thin lines indicate coordination around the Li^+ cation.

crystalline material gave a significantly narrower linewidth than the amorphous phase (Fig. 4a). Samples were also prepared in which there was a mixture of crystalline and amorphous $\text{PEO}_6\text{:LiSbF}_6$ phases present. In these cases two lines were observed, one narrow and one broad, corresponding to the crystalline and amorphous phases respectively. These results demonstrate that the $\text{PEO}_6\text{:LiSbF}_6$ materials used for the conductivity measurements were respectively crystalline and amorphous. This conclusion was reinforced by spin-lattice relaxation time measurements on the spectral lines in Fig. 4a. The signal from the crystalline sample yielded an excellent fit to a single exponential whereas the signal from the amorphous sample required a mixture of at least two exponentials.

The relative contributions of cation and anion transport to conduction in the 6:1 crystalline polymer electrolytes were investigated by NMR spectroscopy. For these studies the $\text{PEO}\text{:LiPF}_6$ system was selected as ^{31}P is a spin- $\frac{1}{2}$ nucleus, unlike ^{121}Sb and ^{75}As , and this makes it easier to follow the anion motion. Figure 4b shows the variation of linewidth over the temperature range 260–310 K for ^7Li and ^{31}P . The ^7Li linewidth narrows significantly whereas the width of the ^{31}P line remains constant with temperature. On decoupling the protons from the phosphorus, the ^{31}P linewidth narrows, indicating that the ^{31}P signal is influenced by interaction with the protons. The lack of narrowing of the ^{31}P signal on heating in the absence of decoupling demonstrates that PF_6^- does not move with respect to the polymer chains. Combining the NMR results with those from the conductivity measurements demonstrates that the ionic conductivity is dominated by Li^+ ; that is, the cation transport number t_+ (the proportion of current carried by the cations) equals 1.

Ion transport in solids involves the hopping of ions between adjacent sites. In the conventional view of ionic conductivity in polymer electrolytes, ions move in a dynamic environment created by the polymer chain motion in the amorphous phase above T_g . A crankshaft-like motion associated with short segments of the polymer chains creates, randomly, suitable coordination sites adjacent to ions so that these ions may hop. Such segmental modes, involving the motion of groups of atoms on the polymer chains, are usually relatively slow, limiting the hopping rate and therefore the maximum conductivity. Despite strenuous efforts over some 20 years involving the preparation of highly amorphous polymer electrolytes with low T_g , the maximum conductivity of such polymer electrolytes remains around $10^{-4} \text{ S cm}^{-1}$ at room temperature. If, on the other hand, the sites to which an ion migrates are already present in the polymer electrolyte owing to its structure, then as soon as the vibrating ion gains sufficient energy to hop, migration will take

place. By implication, designing more organized or ordered polymer electrolytes should enhance conductivity^{21–23}. This may be achieved by forming liquid crystalline polymer electrolytes or by stretching the polymer above T_g (refs 22, 23). But ionic conductivity is expected to be especially favoured in a static environment where permanent pathways for ion transport are present. Crystalline materials have advantages over glasses in this regard, as crystalline ionic conductors may be designed with every pathway being identical and possessing bottlenecks of appropriate size for transport. In contrast, the disordered nature of a glass means that some pathways and bottlenecks will be less favourable.

Further support for the view that ion transport will be favoured in crystalline polymer electrolytes comes from considering ionic conductivity in crystalline ceramic materials such as $\text{Na}\beta\text{Al}_2\text{O}_3$, RbAg_4I_5 , the LISICONS ($\text{Li}_{14}\text{Zn}(\text{GeO}_4)_4$) or $\text{Li}_{0.5}\text{La}_{0.5}\text{TiO}_3$ (refs 4, 24, 25). Some of these materials show ionic conductivities amongst the highest known in the solid state, exceeding by 1 to 3 orders of magnitude the maximum conductivity of conventional amorphous polymer electrolytes. For example, at room temperature RbAg_4I_5 has a conductivity of over $10^{-1} \text{ S cm}^{-1}$, and $\text{Li}_{0.5}\text{La}_{0.5}\text{TiO}_3$ achieves $10^{-3} \text{ S cm}^{-1}$ (refs 4, 24). High ionic conductivity may also be obtained in plastic crystals where ion transport is aided by rotational disorder²⁶.

The level of ionic conductivity in the crystalline 6:1 complex, although significantly higher than the amorphous phase, is not high in general terms. The 6:1 structure clearly promotes Li^+ -ion transport, but it is possible that other structures may yield higher conductivity. Work on the Hollandites (such as $\text{K}_{1.6}\text{Al}_{1.6}\text{Ti}_{6.4}\text{O}_{16}$) shows that ionic conductivity in one-dimensional systems can be severely restricted by just a few impurities or imperfections in the channels²⁷. As a result, crystalline polymers that support ion conduction in two or three dimensions may have higher conductivity than the 6:1 complexes. Ion transport is also favoured when there are substantial numbers of vacancies or interstitials. In the 6:1 complexes, this is not the case. The introduction of

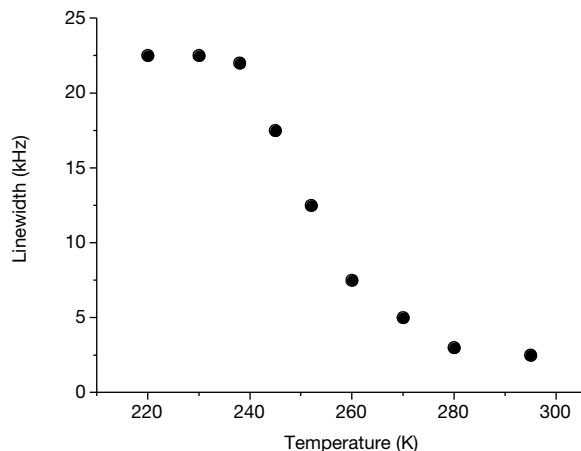


Figure 3 ^1H NMR linewidth as a function of temperature for amorphous $\text{PEO}_6\text{:LiSbF}_6$.

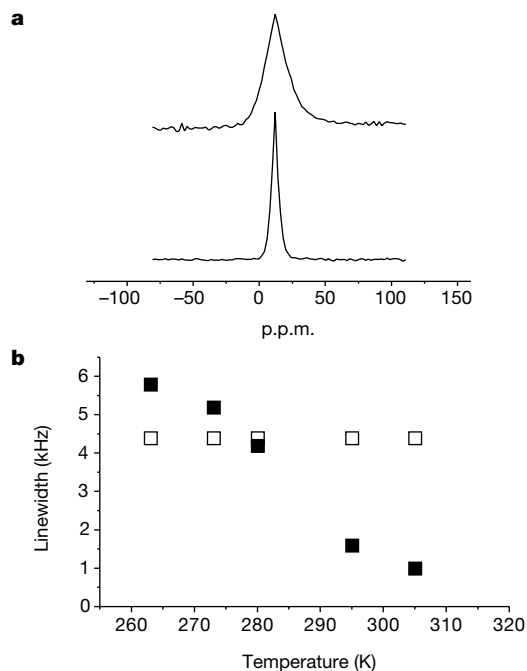


Figure 4 NMR data. These data are on $\text{PEO}_6\text{:LiX}\text{F}_6$ for $\text{X}=\text{Sb}, \text{P}$. **a**, Proton-decoupled static ^7Li spectra for amorphous (upper line) and crystalline (lower line) $\text{PEO}_6\text{:LiSbF}_6$. **b**, Linewidth variations for the crystalline complex $\text{PEO}_6\text{:LiPF}_6$ as a function of temperature: ^{31}P (open squares) and ^7Li (filled squares).

additional defects, even relatively small proportions of extra vacancies or interstitials, into such materials could deliver substantial increases in conductivity, by analogy with the ceramic ionic conductors⁴.

It is important to develop new approaches to the design of polymers with high ionic conductivity²⁸. Our results define a different direction in the search for ionically conducting polymers, one which emphasizes order and structure as important features and challenges us to seek new crystalline polymer electrolytes with suitable structures and with partial occupancy of sites by potentially mobile ions. The selectivity for Li⁺ ion transport in crystalline polymer electrolytes (cation transport number $t_+ = 1$) is a further advantage. □

Methods

Crystalline PEO₆:LiSbF₆ and PEO₆:LiPF₆ were synthesized by dissolving LiSbF₆ or LiPF₆ and PEO ($M_w = 1,000$) in acetonitrile. The PEO was end-capped by methoxy groups. Following removal of the solvent, the resultant material was pressed into a disk for conductivity measurements. The amorphous form of the material was prepared by directly dissolving the salt in molten PEO ($M_w = 100,000$) at high temperatures followed by cooling to room temperature. In both cases the salt and polymer were carefully dried before use, as were the polymer electrolytes after formation. In the case of the crystalline material, infrared spectroscopy was used to confirm that there was no residual acetonitrile. All manipulations were carried out in an argon-filled dry box. Disks of the crystalline and amorphous PEO₆:LiSbF₆ materials were subjected to conductivity measurements by placing them between stainless steel electrodes in a conductivity cell. Different methods were used to establish contact between the polymer and the cell including sputtering or painting silver or platinum metal. No differences were observed between the different methods. Alternating-current impedance measurements were carried out using a Solartron frequency response analyser (1255) and electrochemical interface (1287). A single arc, associated with the response of the polymer electrolyte to the a.c. perturbation, was observed in all cases. The d.c. conductivity was obtained from the low-frequency intercept of this arc with the real axis of the impedance plots. NMR measurements were made on a Bruker MSL 500.

Received 4 April; accepted 21 June 2001.

- Croce, F., Appetecchi, G. B., Persi, L. & Scrosati, B. Nanocomposite polymer electrolytes for lithium batteries. *Nature* **394**, 456–458 (1998).
- Gray, F. M. *Polymer Electrolytes* (RSC Materials Monographs, The Royal Society of Chemistry, Cambridge, 1997).
- Scrosati, B. (ed.) *Applications of Electroactive Polymers* (Chapman & Hall, London, 1993).
- Bruce, P. G. (ed.) *Solid State Electrochemistry* (Cambridge University Press, Cambridge, 1995).
- Fenton, D. E., Parker, J. M. & Wright, P. V. Complexes of alkali metal ions with poly(ethylene oxide). *Polymer* **14**, 589 (1973).
- Armand, M. B., Chabagno, J. M. & Duclot, M. J. in *Fast Ion Transport in Solids* (eds Vashishta, P., Mundy, J. N. & Shenoy, G. K.) 131–136 (North-Holland, Amsterdam, 1979).
- Berthier, C. *et al.* Microscopic investigation of ionic conductivity in alkali metal salts—poly(ethylene oxide) adducts. *Solid State Ionics* **11**, 91–95 (1983).
- Ratner, M. A. in *Polymer Electrolytes Reviews—1* (eds MacCallum, J. R. & Vincent, C. A.) 173–236 (Elsevier Applied Science, London, New York, 1987).
- Druger, S. D., Ratner, M. A. & Nitzan, A. *Solid State Ionics* **9/10**, 1115–1120 (1983).
- Gorecki, W. *et al.* NMR, DSC and conductivity study of the polymer solid electrolytes P(EO)₆(LiC₁₀H₁₇F₃SO₃)₂. *Solid State Ionics* **28–30**, 1018–1022 (1988).
- Shirakawa, H., Louis, E. J., MacDiarmid, A. G., Chiang, C. K. & Heeger, A. J. Synthesis of electrically conducting organic polymers: halogen derivatives of polyacetylene, (CH)_x. *J. Chem. Soc. Chem. Commun.* 578–580 (1977).
- Heeger, A. J. Charge transfer in conducting polymers. *Faraday Discuss. Chem. Soc.* **88**, 203–211 (1989).
- Andreev, Y. G., Lightfoot, P. & Bruce, P. G. A general Monte Carlo approach to structure solution from powder-diffraction data: Application to poly(ethylene oxide)₃:LiN(SO₂CF₃)₂. *J. Appl. Crystallogr.* **18**, 294–305 (1997).
- Andreev, Y. G. & Bruce, P. G. Solving crystal structures of molecular solids without single crystals: a simulated annealing approach. *J. Chem. Soc. Dalton Trans.* 4071–4080 (1998).
- MacGlashan, G. S., Andreev, Y. G. & Bruce, P. G. Structure of the polymer electrolyte poly(ethylene oxide)₃:LiAsF₆. *Nature* **398**, 792–794 (1999).
- Andreev, Y. G., Lightfoot, P. & Bruce, P. G. Structure of the polymer electrolyte poly(ethylene oxide)₃:LiN(SO₂CF₃)₂ determined by powder diffraction using a powerful Monte Carlo approach. *J. Chem. Soc. Chem. Commun.* 2169–2170 (1996).
- Lightfoot, P., Mehta, M. A. & Bruce, P. G. Crystal-structure of the polymer electrolyte poly(ethylene oxide)₃:LiCF₃SO₃. *Science* **262**, 883–885 (1993).
- Gadjourouva, Z., Martin y Marero, D., Andersen, K. H., Andreev, Y. G. & Bruce, P. G. Structures of the polymer electrolyte complexes PEO₆:LiXF₆ (X=P, Sb), determined from neutron powder diffraction data. *Chem. Mater.* **13**, 1282–1285 (2001).
- Roux, C., Gorecki, W., Sanchez, J.-Y., Jeannin, M. & Beloritzky, E. Physical properties of (PPO)₆(LiTFSI) polymer electrolytes: nuclear magnetic resonance investigation and comparison with (PEO)₆(LiTFSI). *J. Phys. Condens. Matter* **8**, 7005–7017 (1996).
- Shi, J. & Vincent, C. A. The effect of molecular weight on cation mobility in polymer electrolytes. *Solid State Ionics* **60**, 11–17 (1993).
- Bruce, P. G. Coordination chemistry in the solid state. *Phil. Trans. R. Soc. Lond. A* **354**, 415–436 (1996).

- Wright, P. V., Zheng, Y., Bhatt, D., Richardson, T. & Ungar, G. Supramolecular order in new polymer electrolytes. *Polym. Int.* **47**, 34–42 (1998).
- Chung, S. H., Wang, Y., Greenbaum, S. G., Golodnitsky, D. & Peled, E. Uniaxial stress effects in poly(ethylene oxide)-LiI polymer electrolyte film. A ⁷Li nuclear magnetic resonance study. *Electrochem. Solid-State Lett.* **2**, 553–555 (1999).
- Inaguma, Y. *et al.* High ionic-conductivity in lithium lanthanum titanate. *Solid State Commun.* **86**, 689–693 (1993).
- Bruce, P. G. & West, A. R. Ion trapping and its effect on the conductivity of LISICON and other solid electrolytes. *J. Solid State Chem.* **53**, 430–434 (1984).
- MacFarlane, D. R., Huang, J. H. & Forsyth, M. Lithium-doped plastic crystal electrolytes exhibiting fast ion conduction for secondary batteries. *Nature* **402**, 792–794 (1999).
- Kudo, T. & Fueki, K. *Solid State Ionics* (Kodansha Ltd, Tokyo; VCH, Weinheim, 1990).
- Angell, C. A., Liu, C. & Sanchez, E. Rubbery solid electrolytes with dominant cationic transport and high ambient conductivity. *Nature* **362**, 137–139 (1993).

Acknowledgements

We thank the UK Engineering and Physical Sciences Research Council for financial support (P.G.B.) and the Overseas Research Scheme for an award (Z.G.).

Correspondence and requests for materials should be addressed to P.G.B. (e-mail: p.g.bruce@st-andrews.ac.uk).

Covariation of carbon dioxide and temperature from the Vostok ice core after deuterium-excess correction

Kurt M. Cuffey* & Françoise Vimeux†

* Department of Geography, and Department of Earth and Planetary Science, 507 McCone Hall, University of California, Berkeley, California 94720-4740, USA

† Laboratoire des Sciences du Climat et de l'Environnement, UMR CEA-CNRS 1572, CEA Saclay, Batiment 709, Orme des Merisiers, 91191 Gif sur Yvette, France

Ice-core measurements of carbon dioxide^{1,2} and the deuterium palaeothermometer reveal significant covariation of temperature and atmospheric CO₂ concentrations throughout the climate cycles of the past ice ages. This covariation provides compelling evidence that CO₂ is an important forcing factor for climate^{3–5}. But this interpretation is challenged by some substantial mismatches of the CO₂ and deuterium records, especially during the onset of the last glaciation, about 120 kyr ago. Here we incorporate measurements of deuterium excess from Vostok^{6,7} in the temperature reconstruction and show that much of the mismatch is an artefact caused by variations of climate in the water vapour source regions. Using a model that corrects for this effect, we derive a new estimate for the covariation of CO₂ and temperature, of $r^2 = 0.89$ for the past 150 kyr and $r^2 = 0.84$ for the period 350–150 kyr ago. Given the complexity of the biogeochemical systems involved, this close relationship strongly supports the importance of carbon dioxide as a forcing factor of climate. Our results also suggest that the mechanisms responsible for the drawdown of CO₂ may be more responsive to temperature than previously thought.

As calculated from Vostok ice-core data¹, the correlation strength of CO₂ and δD for the past 150 kyr is $r^2 = 0.64$, and this is limited primarily by the rapid decrease of δD during and immediately following the last interglacial, 100–130 kyr before present, BP (Fig. 1). If this δD change really reflects a proportional temperature drop, then more than half of the interglacial-to-glacial temperature change occurred before significant removal of atmospheric CO₂, implying a minor contribution of CO₂ to the glacial-age climate forcing. Moreover, such a large early temperature drop is puzzling

because all other known glacial climate forcings were much stronger 70–90 kyr later, during the glacial maximum⁴. The correlation is also limited by CO₂ and δD values at 25 kyr BP relative to those at 75 and 140 kyr BP (Fig. 1).

Oceanic isotopic changes explain some of the δD –CO₂ decorrelation⁸, but current estimates for the oceanic changes⁹ indicate that this correction is minor; the estimated increase (1.2‰) of marine $\delta^{18}O$ implies an essentially 10‰ increase of δD in the source-region vapour, which is decreased by distillation to ~6‰ at Vostok. Earlier solutions^{3,4} to this problem used regression analysis to partition, in an *ad hoc* fashion, the δD history into optimized contributions from a number of possible climate forcings, including greenhouse gases, Southern and Northern Hemisphere insulations (mean annual and monthly), and Northern Hemisphere ice-sheet albedo. These studies attributed 40–60% of Antarctic temperature variations to greenhouse gases. To strengthen this result, a more strongly physical analysis is desirable.

Central to our analysis is the physically based assumption that the temperature difference between vapour sources and precipitation sites is a first-order control on δD of polar precipitation, because net isotopic distillation during atmospheric transport is driven by fractional reduction of air-mass water content¹⁰. Information about source-region climate variations is thus essential for quantitative interpretations of ice-core δD , and, following previous work^{6,11,12} we obtain this information from non-standard covariation of δD and $\delta^{18}O$, defined as the deuterium excess ($d = \delta D - 8\delta^{18}O$; we will refer to this as ‘the excess’). Theory predicts^{11–13} that the excess of polar precipitation varies in response to source-region temperature and relative humidity (see Methods), and that source-region temperature signals can be separated from other influences using marine and ice-core isotopic data⁷. Calibrated versions of this theory successfully explain major features of the global distribution of excess in precipitation¹⁴, including polar snow^{11,12}.

The complete 420-kyr Vostok excess history⁷, spanning four climate cycles, has been measured and analysed. Source temperatures were found to be high during the inceptions of glacial periods, suggesting that the poles were still receiving moisture from warm oceanic sources, even as they cooled in response to feedbacks on isolation changes¹⁵. This would have caused a disproportionately

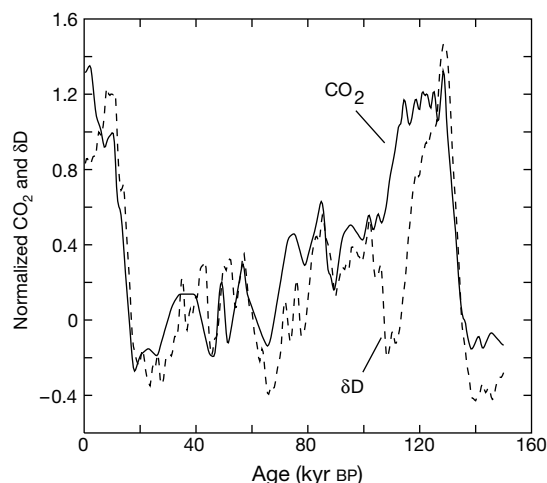


Figure 1 Comparison of carbon dioxide and deuterium histories in Vostok ice³. These are shown normalized to their mean values during the mid-Holocene (5–7 kyr BP) and the last glacial (18–60 kyr BP). Clearly visible are the disproportionately low deuterium values during the mid-glacial (60–80 kyr BP), the glacial inception (95–125 kyr BP), and the penultimate glacial maximum (140–150 kyr BP). δD is defined as $(R/R_s) - 1$, where R is the measured number ratio of HDO to H₂O, and R_s is the standard ratio (Vienna SMOW).

large drop in δD of Antarctic precipitation at this time. To address such effects, we use the following framework.

For the relatively small-magnitude isotopic changes associated with climate changes at Vostok, isotope model results can be approximated⁷ as linear relations. These relations ascribe changes of Antarctic precipitation's δD ($\Delta\delta_A$) and excess (Δd) to changes in local Antarctic temperature (ΔT_A), effective source temperature (ΔT_s), and marine isotopic composition ($\Delta\delta_m$, measured as changes of $\delta^{18}O$). We therefore specify that

$$\Delta\delta_A = \gamma_A \Delta T_A - \gamma_s \Delta T_s + \gamma_m \Delta\delta_m \quad \gamma_A, \gamma_s, \gamma_m > 0 \quad (1)$$

and

$$\Delta d = \beta_s \Delta T_s - \beta_A \Delta T_A - \beta_m \Delta\delta_m \quad \beta_s, \beta_A, \beta_m > 0 \quad (2)$$

Table 1 gives model-based estimates for the sensitivity coefficients. The coefficient β_s describes the important thermometric property of the excess that in principle allows a correction to be made for the ΔT_s term in equation (1). Predicted values for β_s are 1.26‰ °C⁻¹, from a single-trajectory Rayleigh-type model⁷, and 1.06‰ °C⁻¹ from simulations with the GISS general circulation model¹⁶ (GCM). Relative humidity variations are included in these values by using the observed modern correlation of sea surface temperature and relative humidity, which GCM simulations suggest persisted in glacial climates^{7,17}.

Given the two measured histories $\Delta\delta_A$ and Δd , the estimated history $\Delta\delta_m$ (see Supplementary Information), and the model-based sensitivities, equations (1) and (2) determine a unique ΔT_A history. To interpret this ΔT_A as a response to known climate forcings which are global or hemispheric in scale, we distinguish between the Antarctic changes ΔT_A (of which Vostok is assumed to

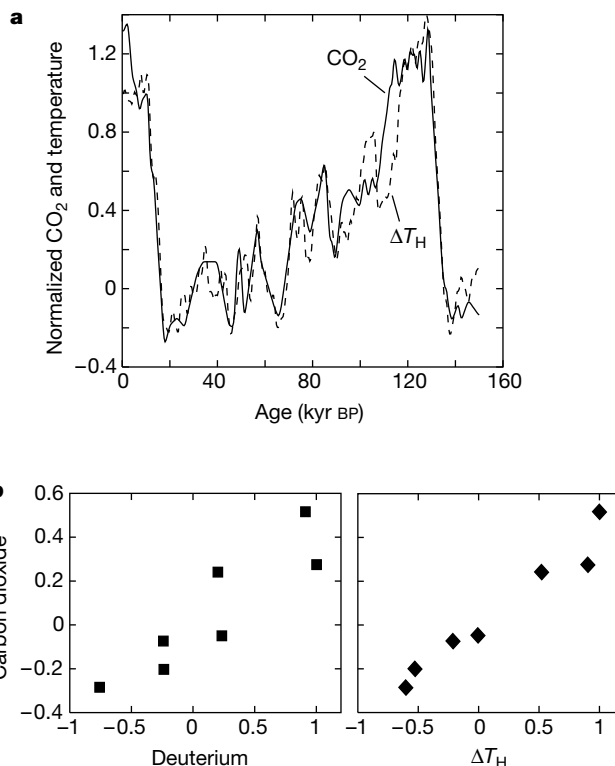


Figure 2 Comparison of carbon dioxide and reconstructed Southern Hemisphere temperature, ΔT_H . **a**, Their histories, normalized as in Fig. 1, using $\beta_s = 1.1$. Results are nearly identical using β_s from GCM or Rayleigh models (for the GISS value, $\beta_s = 1.06$ and $r^2 = 0.886$, while for the Rayleigh value, $\beta_s = 1.26$ and $r^2 = 0.876$). **b**, Comparisons of means for seven specific intervals selected to highlight the difference (in units of kyr BP: 0–6, 6–25, 25–60, 60–90, 90–125 and 135–150).

be representative) and average temperature changes over a broader region of the Southern Hemisphere ΔT_H . The latter are defined conceptually as a weighted average of Antarctic changes and lower-latitude changes ΔT_l using a geometric constant f_l (given the surface area of Antarctica relative to that for the source regions, $f_l \approx 0.7-0.9$) so that $\Delta T_H = f_l \Delta T_l + (1 - f_l) \Delta T_A$.

Palaeoclimate studies and climate models both show that feedbacks cause polar amplification of hemispheric climate changes^{18,19}. Thus the meridional temperature gradient $T_l - T_A$ increases as the hemisphere cools. In addition, there are changes $G(t)$ in the meridional temperature gradient due to changes in the meridional gradient of mean annual insolation; these are adopted without alteration from a GCM²⁰ (see Methods) and are of small magnitude. Other changes in the temperature gradient may have occurred but are not known. The most parsimonious statement is thus:

$$\Delta T_l - \Delta T_A = -\epsilon_g \Delta T_H + G(t) \quad \epsilon_g > 0 \quad (3)$$

where ϵ_g is an amplification constant. Excepting a scaling constant (k_*) that is irrelevant to the present analysis because it does not affect correlations, the calculation of ΔT_H uses only the measured $\Delta \delta_A$ data, the excess-based ΔT_s correction, the marine isotopic correction, and the model-based $G(t)$, and therefore contains no *ad hoc* terms (see Supplementary Information). The relation is:

$$\Delta T_H = k_*^{-1} [\Delta \delta_A + c_* \Delta d + (c_* \beta_m - \gamma_m) \Delta \delta_m + f_l (\gamma_A - \beta_A c_*) G(t)] \quad (4)$$

$$c_* = \gamma_s \beta_s^{-1} \quad (5)$$

$$k_* = (1 + f_l \epsilon_g) (\gamma_A - \beta_A c_*) \quad (6)$$

A fundamental question is whether these 'hemispheric' temperature changes can be explained by the sum of known climate forcings $\Delta \Phi$ (in W m^{-2}) so that $\Delta T_H = \epsilon_f \Delta \Phi$ (see Methods), where ϵ_f is a net feedback factor.

Adopting either predicted value for β_s yields a remarkably strong relationship between the greenhouse-gas histories and this estimate of Southern Hemisphere temperature (Fig. 2). The correlation of ΔT_H is only significantly affected by β_s and γ_s (Table 1) whose values are taken from physical models, yet the covariance of ΔT_H and CO_2 is very high ($r^2 = 0.886$), and the covariance of ΔT_H with the combined CO_2 and CH_4 forcings is 0.90. Of this, 0.165 is due to the excess correction for ΔT_s , 0.07 is due to the oceanic composition correction, and 0.015 is due to the $G(t)$.

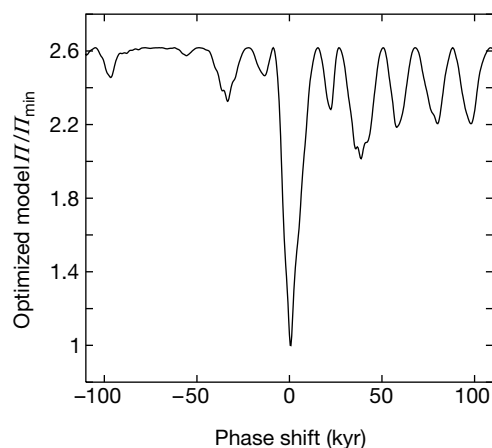


Figure 3 Effect of phase-shifting the deuterium-excess data. Shown is the performance for the best possible model achieved if the deuterium-excess data are uniformly phase-shifted by the amount specified. The strong minimum near zero phase-shift is solid evidence that the data are meaningful. Π is defined in the text; $\Pi_{\min}$ is Π for the optimal model $c_* = 3.75$.

The reconstructed ΔT_H can be compared rigorously to known climate forcings over a time interval τ by defining a mean square error performance index as $\Pi = \int_{\tau} [\Delta T_H - \epsilon_f \Delta \Phi]^2 dt$ (see Methods). The best possible reconstruction for given data (an optimal model) is one which minimizes this index. For the most recent 150 kyr, setting $\beta_s = 1.18$ (the average of the predicted values (Table 1) reduces Π by 59% relative to using $\beta_s = 0$, allowing only a scaling coefficient k_* to be optimized for each. The excess correction thus directly yields a temperature history that more closely matches climate forcings.

More generally, we can examine the role of the excess correction by letting both c_* and k_* be free parameters in the minimization of Π . This yields an optimal model with $c_* = 3.75$ or 3.50 (excluding or including albedo forcing from $\Delta \Phi$, respectively), for which $\beta_s = 0.99$ or 1.06‰ °C^{-1} . Monte Carlo analysis yields a 95% confidence interval on β_s of (0.73, 1.59) (see Methods). The accordance of these optimal β_s values with isotope model predictions shows that our interpretations are internally consistent, and suggests a method for demonstrating that the excess correction is physically meaningful, as follows.

We evaluate whether it is possible for the improved covariation due to the excess data to arise by chance. Specifically, given any random curve with the same power spectrum as the measured excess history, how likely is it that our method will find an optimal solution (values for c_* and k_*) that performs as well as, or better than, the one based on data? Using a Monte Carlo approach and fake 'excess-like' histories constructed by randomizing the phases of the spectral components of the measured excess, we show that the probability of this error is negligible (see Methods). This is graphically demonstrated for the case of a uniform phase-shift (Fig. 3). This strongly suggests that the excess history contains meaningful climate information.

Another demonstration of the excess data's predictive power results from applying the same analysis to the older part of the Vostok record (150–350 kyr), an independent data set which spans two more interglacial periods, and for which excess has recently been measured⁷. Applying $\beta_s = 1.1$ to this period also significantly improves the calculation, reducing Π by 38% relative to the case of $\beta_s = 0$, and giving $r^2 = 0.84$ and 0.86 (for CO_2 alone, and for combined CO_2 and CH_4 , respectively), with no adjustable parameters. Optimization yields a best value of $\beta_s = 1.07$ or 1.33 (excluding and including the albedo forcing, respectively), nearly equivalent to that for 0–150 kyr BP. This coherence further supports our methodology.

Our results give strength to the conclusion^{4,18} that CO_2 is an important climate forcing on the modern Earth, irrespective of whether other factors are more important on very long geologic timescales²¹. Further, our results strengthen the hypothesis⁵ that the

Table 1 Values for isotopic model parameters

Parameter	Value	Units	Δr^2	Δr^2
			for 20% larger	for 20% smaller
γ_A	7.1*	‰ °C ⁻¹	+0.2	-0.4
γ_s	3.7†	‰ °C ⁻¹	+1.2	-2.4
γ_m	4.8‡	Dimensionless	+0.3	-0.4
β_A	0.5§	‰ °C ⁻¹	-0.07	+0.06
β_s	1.06	‰ °C ⁻¹	-1.9	+1.3
β_m	1.26¶	‰ °C ⁻¹	-3.6	+0.66
β_m	2.75#	Dimensionless	-0.8	+0.6

Values shown are adopted without modification from ref. 7. Also shown is the sensitivity of the ΔT_H - CO_2 correlation to each model parameter. The sensitivity is given as the percentage change in the r^2 value for the correlation resulting from 20% reductions and enhancements of the parameters ($r^2 = 88.6\%$ for the standard parameter values).

* Rayleigh models⁷, and spatial gradient¹⁴.

† Rayleigh models³⁰. GCMs yield similar values³⁰.

‡ Rayleigh model. This number must be close to accurate because it simply measures the total deuterium isotopic depletion between the source region and Vostok (~44%).

§ Rayleigh model⁷.

|| GISS model^{7,18}. Includes covariation of sea surface temperature and relative humidity.

¶ Rayleigh model⁷, including covariation as in GISS model above. Also similar to ref. 12.

Rayleigh model⁷. GISS model yields similar value, 2.9 (ref. 16).

long-term synchrony of glacial–interglacial cycling between Northern and Southern Hemispheres is due to greenhouse-gas variations, and feedbacks associated with them. In particular, we have shown that most Southern Hemisphere ΔT can be explained, in the correlative sense, without recourse to any Northern Hemisphere forcings, including insolation at 65° N and ice-sheet albedo.

A further result relevant to carbon-cycle studies is that much of the very long delay (10–15 kyr) and disproportionate magnitude of the CO₂ decrease relative to the δD decrease during the last glacial inception is an artefact. The actual delay between CO₂ decrease and temperature change was probably not more than 5 kyr: such a delay is much more easily explained in terms of the time required for ocean circulation response and deep-water ventilation. Nonetheless, the correspondence of CO₂ and temperature during the glacial inception was not as strong as during the last deglacial transition²². □

Methods

Source-region isotope effect

The composition of evaporation for the isotope j is calculated as^{11,13}

$$1 + \delta_j = \frac{1}{\alpha_j} \left(\frac{1 - k_j}{1 - k_j h} \right) (1 + \delta_{jm}) \quad (7)$$

where h is relative humidity, k_j is a constant, α_j the equilibrium fractionation factor, and δ_{jm} the marine composition. The temperature dependence arises almost entirely from the temperature dependence of the α_j . Such source-region changes are further modified during atmospheric transport, necessitating the use of atmospheric isotopic models to estimate the coefficients in Table 1.

Marine isotopic history

For the past 150 kyr we estimate $\Delta \delta_m$ directly from radiometrically dated sea-level history^{23–27} and calibrate this to the Last Glacial Maximum ocean composition inferred from pore water⁹ (see Supplementary Information). For the older record (150–350 kyr BP), sea-level data of this quality are not available, and we instead use the foram $\delta^{18}O$ history²⁸, which is less desirable because it includes temperature effects. The GT4 timescale is used for the Vostok data⁴.

Insolation gradient term

The mean annual meridional insolation gradient varies with the 41-kyr obliquity cycle²⁹. Excess data illustrate⁶ a real climatic effect of this gradient. Thus we adopt the simplest possible assumption, that $G(t) = \gamma_l I(t)$, where I is the obliquity insolation gradient and γ_l a constant sensitivity ($\gamma_l = 0.2^\circ C m^{-2} W^{-1}$), whose value we estimate directly from Holocene atmosphere–ocean coupled GCM simulations³⁰ using the modelled change in meridional sea surface temperature gradient per change in insolation, and assuming that the temperature effect in the interior of the Antarctic continent can be estimated as the sea surface temperature changes at 60° S. Fortunately, the results of our analysis depend only weakly on the $G(t)$ term. Zeroing and doubling the term with $G(t)$ varies the r^2 for ΔT_H –CO₂ correlation from 0.87 to 0.89, and changes the optimal β_5 to 0.87 and 1.28, respectively.

Summed climate forcing

$\Delta \Phi = \Delta \Phi_{CO_2} + a_{CH_4} \Delta \Phi_{CH_4} + a_a \Delta \Phi_a$, where the two gas terms are the Vostok concentration histories² normalized to a range of one, $a_{CH_4} = 0.15$, and $\Delta \Phi_a$ is the albedo forcing, calculated using the sea-level history normalized to a range of one. $a_a = 1$ for tests including the albedo forcing¹⁸ and $a_a = 0$ for gas-only scenarios. All other forcings (such as water vapour, N₂O, aerosol, and clouds) are assumed to be feedbacks, because data or theory are insufficient for calculation of radiative forcing histories.

Optimization of Π

The scaling constant is $\hat{k}_* = \epsilon_l k_*$. Singular value decomposition is used to optimize the equivalent expression:

$$\int_t [\Delta \delta_a - \gamma_m \Delta \delta_m + c_*(\Delta d + \beta_m \Delta \delta_m) - \hat{k}_* \Delta \Phi + f_l(\gamma_a - \beta_a c_*) G(t)]^2 dt \quad (8)$$

Confidence limits on c_*

Optimal c_* is recalculated 1,000 times using random perturbations of the primary uncertain inputs, and its distribution compiled. Perturbed quantities are: (1) the coefficient of the $G(t)$ term, treated as a uniform random variable distributed between 0.5 and 2 times its nominal value; (2) the relative timing of sea-level change and Vostok data, which is stretched by an amount that grows linearly with time and by 125 kyr BP reaches a random number distributed uniformly between –5 and +5 kyr; (3) the residual of ΔT_H and $\Delta \Phi$ (which is assumed to represent an unknown set of physical processes), which is

decomposed into spectral components whose phases are randomly shifted before reconstruction in the time domain to give an extra function that is added to the definition of Π ; (4) the gas ages are shifted relative to the isotope ages by an amount that is uniformly distributed between ± 0.5 kyr.

Randomized-phase analysis

The excess data are decomposed into spectral components, which are each independently randomly phase-shifted by a number between π and $-\pi$. These are inverse-transformed to generate the fake excess-like history, which is treated in the analysis in the same way as the Δd data. The distribution of Π for 1,000 of these histories indicates that there is a probability of less than 0.02 that the actual improvement of Π from the excess data could arise by chance, and a probability of less than 0.07 that more than half of this improvement in Π could arise by chance.

Received 7 February; accepted 19 June 2001.

- Petit, J. R. *et al.* 420,000 years of climate and atmospheric history revealed by the Vostok deep Antarctic ice core. *Nature* **399**, 429–436 (1999).
- Pepin, L., Raynaud, D., Barnola, J. M. & Loutre, M. F. Hemispheric roles of climate forcings during glacial–interglacial transitions as deduced from the Vostok record and LLN-2D model experiments. *J. Geophys. Res.* (in the press).
- Genthon, C. *et al.* Vostok ice core: climatic response to CO₂ and orbital forcing changes over the last climatic cycle. *Nature* **329**, 414–418 (1987).
- Lorius, C., Jouzel, J., Raynaud, D., Hansen, J. & LeTreut, H. Greenhouse warming, climate sensitivity and ice core data. *Nature* **347**, 139–145 (1990).
- Barnola, J. M., Raynaud, D., Korotkevich, Y. S. & Lorius, C. Vostok ice core provides 160,000-year record of atmospheric CO₂. *Nature* **329**, 408–414 (1987).
- Vimeux, F., Masson, V., Jouzel, J., Stievenard, M. & Petit, J. R. Glacial–interglacial changes in ocean surface conditions in the Southern Hemisphere. *Nature* **398**, 410–413 (1999).
- Vimeux, F. *et al.* A 420,000 year deuterium excess record from East Antarctica: Information on past changes in the origin of precipitation at Vostok. *J. Geophys. Res.* (in the press).
- Jouzel, J. *et al.* Extending the Vostok ice-core record of paleoclimate to the penultimate glacial period. *Nature* **364**, 407–412 (1993).
- Schrag, D. P., Haupt, G. & Murray, D. W. Pore fluid constraints on the temperature and oxygen isotopic composition of the glacial ocean. *Science* **272**, 1930–1932 (1996).
- Dansgaard, W. Stable isotopes in precipitation. *Tellus* **16**, 436–447 (1964).
- Johnsen, S. J., Dansgaard, W. & White, J. W. C. The origin of Arctic precipitation under present and glacial conditions. *Tellus* **41**, 452–468 (1989).
- Petit, J. R., White, J. W. C., Young, N. W., Jouzel, J. & Korotkevich, Y. S. Deuterium excess in recent Antarctic snow. *J. Geophys. Res.* **D 96**, 5113–5122 (1991).
- Merlivat, L. & Jouzel, J. Global climatic interpretation of the deuterium–oxygen 18 relationship for precipitation. *J. Geophys. Res.* **C 84**, 5029–5033 (1979).
- Jouzel, J. *et al.* Simulations of the HDO and H₂¹⁸O atmospheric cycles using the NASA/GISS General Circulation Model: The seasonal cycle for present day conditions. *J. Geophys. Res.* **92**, 14739–14760 (1987).
- Khodri, M. *et al.* Simulating the amplification of orbital forcing by ocean feedbacks in the last glaciation. *Nature* **410**, 570–574 (2001).
- Delaygue, G. *Relations entre Surface Océanique et Composition Isotopique des Précipitations Antarctiques: Simulations pour Différents Climats*. Thesis, Univ. Aix-Marseille 3 (2000).
- Vimeux, F. *et al.* Holocene hydrological cycle changes in the Southern Hemisphere documented in east Antarctic deuterium excess records. *Clim. Dyn.* **17**, 503–513 (2000).
- Hansen, J. *et al.* in *Climate Processes and Climate Sensitivity* (eds Hansen, J. E. & Takahashi, T.) 130–163 (Geophysical Monograph 29, American Geophysical Union, Washington DC, 1984).
- Cuffey, K. *et al.* Large Arctic temperature change at the Wisconsin–Holocene glacial transition. *Science* **270**, 455–458 (1995).
- Braconnot, P., Marti, O. & Joussaume, S. Adjustments and feedbacks in a global coupled ocean–atmosphere model. *Clim. Dyn.* **13**, 507–519 (1997).
- Veizer, J., Goddard, Y. & Francois, L. M. Evidence for decoupling of atmospheric CO₂ and global climate during the Phanerozoic eon. *Nature* **408**, 698–701 (2000).
- Monnin, E. *et al.* Atmospheric CO₂ concentrations over the last glacial termination. *Science* **291**, 112–114 (2001).
- Bard, E., Hamelin, B., Fairbanks, R. G. & Zindler, A. Calibration of the ¹⁴C timescale over the past 30,000 years using mass-spectrometric U–Th ages from Barbados corals. *Nature* **345**, 405–410 (1990).
- Esat, T. M., McCulloch, M. T., Chappell, J., Pillans, B. & Omura, A. Rapid fluctuations in sea level recorded at Huon Peninsula during the penultimate deglaciation. *Science* **283**, 197–201 (1999).
- Stirling, C. H., Esat, T. M., Lambeck, K. & McCulloch, M. T. Timing and duration of the last interglacial: evidence for a restricted interval of widespread coral reef growth. *Earth Planet. Sci. Lett.* **160**, 745–762 (1998).
- Chappell, J. *et al.* Reconciliation of late Quaternary sea levels derived from coral terraces at Huon Peninsula with deep sea oxygen isotope records. *Earth Planet. Sci. Lett.* **141**, 227–236 (1996).
- Fairbanks, R. G. A 17,000-year glacio-eustatic sea level record: influence of glacial melting rates on Younger Dryas event and deep-ocean circulation. *Nature* **342**, 637–642 (1989).
- Bassiot, F. C. *et al.* The astronomical theory of climate and the age of the Brunhes–Matuyama magnetic reversal. *Earth Planet. Sci. Lett.* **126**, 91–108 (1994).
- Loutre, M. F. *Paramètres Orbitaux et Cycles Diurne et Saisonnier des Insolutions*. Thesis, Univ. Catholique de Louvain (1993).
- Jouzel, J. *et al.* Validity of the temperature reconstruction from water isotopes in ice cores. *J. Geophys. Res.* **C 102**, 26471–26487 (1997).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank J. R. Petit *et al.* for access to the data of ref. 1, and I. Fung and the staff of the

Berkeley Atmospheric Sciences Center, J. Jouzel, V. Masson, J. Kirchner and R. Alley for discussions. This work was supported by the US NSF; the Vostok project was supported in France by the PNEDC.

Correspondence and requests for materials should be addressed to K.M.C. (e-mail: kcuffey@socrates.berkeley.edu).

Chemical interaction of Fe and Al_2O_3 as a source of heterogeneity at the Earth's core–mantle boundary

L. Dubrovinsky*, H. Annersten*, N. Dubrovinskaia*, F. Westman*, H. Harryson*, O. Fabrichnaya† & S. Carlson‡

* Department of Earth Sciences, Uppsala University, S-753 36 Uppsala, Sweden

† Max Planck Institut für Metallforschung, Stuttgart D 70569, Germany

‡ European Synchrotron Radiation Facility, Grenoble 38043, France

Seismological studies have revealed that a complex texture or heterogeneity exists in the Earth's inner core and at the boundary between core and mantle^{1–4}. These studies highlight the importance of understanding the properties of iron when modelling the composition and dynamics of the core and the interaction of the core with the lowermost mantle^{5–7}. One of the main problems in inferring the composition of the lowermost mantle is our lack of knowledge of the high-pressure and high-temperature chemical reactions that occur between iron and the complex Mg–Fe–Si–Al oxides which are thought to form the bulk of the Earth's lower mantle. A number of studies^{6,8–12} have demonstrated that iron can react with MgSiO_3 -perovskite at high pressures and high temperatures, and it was proposed^{6,8} that the chemical nature of this process involves the reduction of silicon by the more electropositive iron. Here we present a study of the interaction between iron and corundum (Al_2O_3) in electrically- and laser-heated diamond anvil cells at 2,000–2,200 K and pressures up to 70 GPa, simulating conditions in the Earth's deep interior. We found that at pressures above 60 GPa and temperatures of 2,200 K, iron and corundum react to form iron oxide and an iron–aluminium

alloy. Our results demonstrate that iron is able to reduce aluminium out of oxides at core–mantle boundary conditions, which could provide an additional source of light elements in the Earth's core and produce significant heterogeneity at the core–mantle boundary.

According to arguments based on cosmic abundance, Al_2O_3 is likely to be the next most abundant component in the Earth's lower mantle after MgO , FeO and SiO_2 (refs 13, 14). Moreover, the D'' layer may be enriched in refractories such as Al_2O_3 and CaO (refs 13, 15). Therefore, possible chemical reactions in the Fe– Al_2O_3 system can provide an important model for processes at the core–mantle boundary (CMB). It was demonstrated recently that even small amounts of Al can markedly change the relative proportion of Fe^{3+} in $(\text{Mg,Fe})(\text{Si,Al})\text{O}_3$ perovskite^{16,17}. Consequently, chemistry of the Fe–Al–O system is important for the whole of the Earth's mantle. Knowledge of exchange reactions at high pressures and high temperatures between a metal from one side and refractory oxides and silicates from another side is important for understanding the differentiation of the early Earth^{18,19}. Therefore, there are a number of reasons to study interactions between aluminium oxide and iron in the megabar pressure range and at high temperatures.

At ambient pressure corundum and iron do not react. Reliable thermodynamic data for the Fe–Al–O system exist²⁰ at pressures below 20 GPa. We made calculations of phase equilibria in the Fe–Al–O system to check if Al_2O_3 interacted with metallic iron at subsolidus conditions. Principal phases were Al_2O_3 (corundum), Fe_xO (wüstite), FeAl_2O_4 (hercynite), Fe_2O_3 (haematite), Fe_3O_4 (magnetite) and Fe–Al (f.c.c.) solid solution. Calculations for the composition $\text{Al}_2\text{O}_3 + \text{Fe}$ at temperatures up to 2,000 K and pressures up to 20 GPa show that corundum (Al_2O_3) is in equilibrium with pure metallic iron (f.c.c.), and no reaction proceeds under these conditions.

There are indications that such a reaction is possible at much higher pressures^{21,22}, but no systematic investigations of the interaction between iron and aluminium oxide at different pressures and temperatures so far been made. We therefore decided to conduct *in situ* and *ex situ* studies of the possible reactions between Fe and Al_2O_3 in heated diamond anvil cells by means of X-ray powder diffraction and Mössbauer spectroscopic analysis.

Experiments were done at Uppsala Laboratory in Sweden and on the beamline ID 30 at European Synchrotron Radiation Facility (Grenoble, France). At Uppsala we obtained powder X-ray diffraction

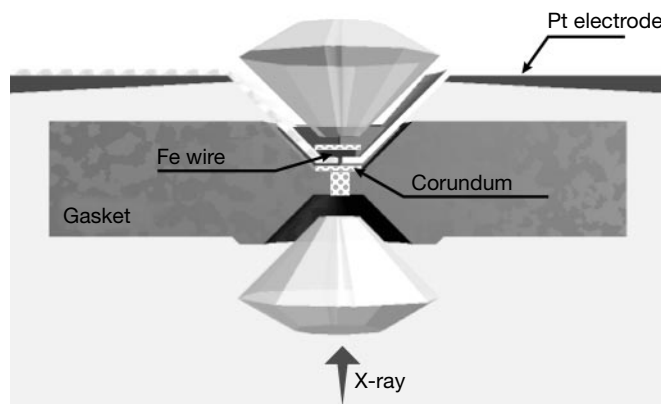


Figure 1 Schematic diagram of electrical heating assemblage. A rhenium gasket of 250 μm thickness was indented to 25–30 μm between diamonds with 300 μm horizontal faces, and a hole 100–110 μm in diameter was drilled. The gasket was covered by corundum-based cement and pure corundum was placed in the hole and

around the wire. A platinum wire of 0.2 mm diameter flattened to a thickness less than 10 μm was used as electrical leads. The iron wire was heated by direct current with a stabilized power supply operated at the 18 V/20 A range.

data on a Siemens X-ray system consisting of a Smart CCD (charge-coupled device) area detector and a direct-drive rotating anode as X-ray generator (18 kW). The MoK α radiation (tube voltage 50 kV, tube current 24 mA, cathode gun 0.1 $\times$ 1 mm) is focused with a capillary X-ray optical system to a diameter of 40 μ m FWHM²³. At ESRF we collected powder diffraction data using a fine incident X-ray beam (approximately rectangular shape with dimensions 12 $\times$ 15 μ m²) of 0.3738 Å wavelength on the MAR345 imaging plate²⁴. The collected images were integrated to obtain conventional diffraction spectra.

We carried out Mössbauer experiments using a conventional transmission geometry²⁴. Measurements were made on temperature-quenched samples mounted in diamond anvil cells, with the γ -rays led through the gasket hole. A point source of ⁵⁷Co in rhenium, nominal value 60 mCi, was used in the experiment and calibrated against α -Fe at room temperature. As starting materials, we used pure (99.999%) Al₂O₃ (lattice parameters are $a = 4.7581(3)$ Å, $c = 12.99(1)$ Å) and α -iron ($a = 2.8662(1)$ Å) (wire or foil). Great care was taken to eliminate contamination by oxygen or water. Before being loaded into the cell, corundum was dried for at least 30 hours at 420–450 K. In all experiments we loaded and closed cells in an inert atmosphere (He or Ar). Pressures were determined at ambient temperature from ruby fluorescence and temperatures were measured spectroradiometrically²⁵.

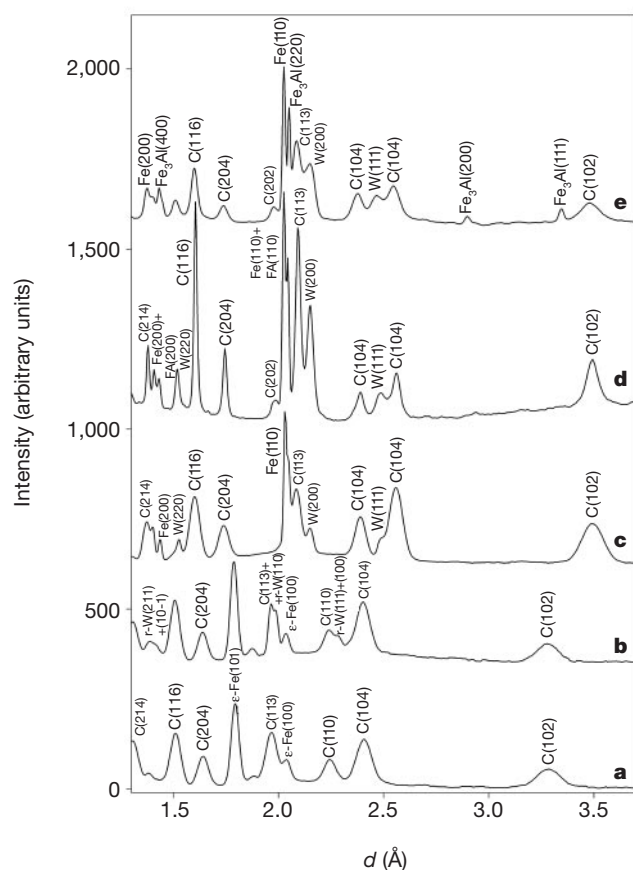


Figure 2 Examples of X-ray diffraction patterns collected after heating of iron and corundum at different pressures and temperatures. **a**, 63(3) GPa and 1,400(50) K; **b**, 65(3) GPa and 2,200(50) K; **c**, completely decompressed sample heated at 65(3) GPa and 2,200(50) K; **d**, and **e**, different spots of the sample heated at 56(2) GPa and 2,000(150) K. Diffraction patterns **a** to **c** are of electrically heated samples collected at the laboratory X-ray source. Diffraction patterns **d** and **e** are of laser-heated samples collected at the ID30 beam line at ESRF. (C, corundum; Fe, α -Fe; W, wüstite; r-W, rhombohedral high-pressure modification of FeO; FA, iron–aluminium alloy.)

In the first experiment, a thin iron wire (5–7 μ m in diameter) was heated electrically (Fig. 1)^{26,27}. For each of the three different loadings, rhenium gaskets were indented between two diamonds, and a hole of 100–110 μ m in diameter was drilled. The gasket was covered by corundum-based cement and pure corundum was placed into the hole and around the wire (Fig. 1). A platinum wire was used as electrical leads. The thin iron wire was heated by direct current. Samples were heated during 40 to 60 min at 22(1) GPa and 1,850(50) K; 34(2) GPa and 2,050(50) K; 38(2) GPa and 2,050(50) K; 45(2) GPa and 1,700(50) K; 46(2) GPa and 2,100(50) K; 55(3) GPa and 1,450(40) K; 54(3) GPa and 2,200(50) K; 63(3) GPa and 1,400(50) K; 65(3) GPa and 220(50) K.

Samples were studied before and after heating with X-rays at the Uppsala Laboratory. At all pressures and temperatures up to 63(3) GPa and 1,400(50) K, samples contained only a mixture of ϵ -Fe and corundum (Fig. 2a; see also Supplementary Information) after heating. But at 65(3) GPa after heating at 2,200(50) K, the

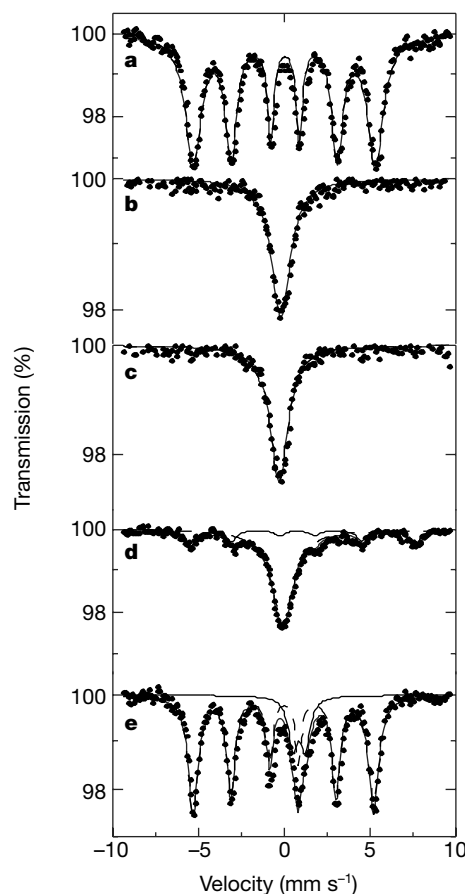


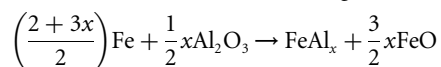
Figure 3 Mössbauer spectra obtained at room temperature and different pressures before and after laser heat treatment at 2,000(150) K. **a**, at 3.6(1) GPa; **b**, at 29 to 30 GPa after heating; **c**, at 56 to 60 GPa before heating; **d**, at 56 to 60 GPa after heating; **e**, on decompression to ambient pressure at pressures above 11 GPa the spectrum contains the absorption pattern from the paramagnetic ϵ -Fe. This pattern is the only pattern observed after heating at 29 to 30 GPa but a second pattern is clearly recognized after heating at 56 to 60 GPa. This second magnetic pattern has identical Mössbauer parameters to those reported²⁹ for rhombohedral FeO—hyperfine field $H_{\text{hyp}} = 40(1)$ T and isomer shift $IS = 0.89(2)$ mm s⁻¹ compared with $H_{\text{hyp}} = 42(1)$ T and $IS = 0.87(1)$ relative to α -Fe as given by Pasternak *et al.*²⁹. This hyperfine pattern could be observed down to 30 GPa, together with the absorption line from ϵ -Fe, on decompressing the diamond anvil cell. At ambient pressure the spectrum shows the well-known six-line pattern from α -Fe and the paramagnetic doublet of cubic FeO with Mössbauer parameters $IS = 0.85(1)$ mm s⁻¹ and quadrupole splitting $0.71(1)$ mm s⁻¹.

diffraction pattern showed several rather weak additional reflections (Fig. 2b; see also Supplementary Information). All these reflections could be interpreted as belonging to the rhombohedral FeO phase²⁴. After decompression, the samples contained diffraction lines of α -Fe ($a = 2.871(2) \text{ \AA}$), corundum ($a = 4.7583(4) \text{ \AA}$, $c = 12.990(2) \text{ \AA}$) and cubic wüstite ($a = 4.321(2) \text{ \AA}$) (Fig. 2c). Although the lattice parameters of corundum did not change after the experiment, the lattice parameter of iron increased. Moreover, the reflection (110) of α -Fe was slightly asymmetric from the side of higher d spacing (Fig. 2c). Such changes in the diffraction pattern of iron correspond to the formation of a Fe–Al alloy with about 2 wt% Al (ref. 28).

In the second experiment, a thin foil (7 to 9 μm) of ^{57}Fe metal was sandwiched between layers of pure corundum in a hole of 100–110 μm in diameter drilled in a rhenium gasket. In the two different loadings, samples were pressurized in the diamond anvil cell using diamonds with 350 μm faces and heated from both sides by using a Nd:YAG laser to a temperature of 2,000(150) K (ref. 25). The laser beam-size was 20–25 μm . The surface of the sample was slowly scanned in 20- μm steps and each point was heated for at least 15 min. Before and after heating Mössbauer spectra were collected through the cell (Fig. 3). Mössbauer spectroscopy sees only the iron-containing phases. At low pressure only the spectrum of α -Fe is present, and shows some line broadening due to saturation effects from the enrichment of ^{57}Fe in the sample (95% ^{57}Fe). At pressures above 11 GPa the spectrum contains the absorption pattern from paramagnetic ϵ -Fe. This is the only pattern observed after heating at 29 GPa, but a second pattern is seen after heating at 56–60 GPa (Fig. 3). This second magnetic pattern has Mössbauer parameters identical to those reported for rhombohedral FeO by Pasternak *et al.*²⁹. On decompression, this pattern, together with the absorption line from ϵ -Fe, could be observed down to 30 GPa. At ambient pressure the spectrum shows the well-known six-line pattern from α -Fe and the paramagnetic doublet of cubic FeO.

There is a noticeable drop of the isomer shift values for the ϵ -Fe single absorption line after heating at 56 GPa. The isomer shift dropped from -0.30 mm s^{-1} before heating to -0.16 mm s^{-1} after heating, at the same time as the rhombohedral FeO appeared in the spectrum. This cannot be due to pressure decrease after heating (the pressure remained the same), but could be explained by a change in the surroundings of the iron nucleus, for instance a substitution of Al into the ϵ -phase. Samples recovered from the gasket out of the cell were studied at ESRF. X-ray diffraction spectra were collected at different points of the sample following a mesh with 25- μm step size. High-resolution synchrotron X-ray data allows the diffraction peaks from pure non-reacted iron ($a = 2.8660(2) \text{ \AA}$) to be resolved from those of reacted iron alloyed with 3 wt% Al²⁸ ($a = 2.8723(2) \text{ \AA}$) (Fig. 2d). Moreover, in one of the spots we observed additional reflections at 3.346 $\text{\AA}$ and 2.897 $\text{\AA}$ which belong to cubic Fe₃Al ($a = 5.7946(4) \text{ \AA}$) (Fig. 2e; see also Supplementary Information). All diffraction patterns also contain reflections from corundum ($a = 4.7581(5) \text{ \AA}$, $c = 12.991(3) \text{ \AA}$) that was unchanged during the experiment, and from wüstite ($a = 4.318(3) \text{ \AA}$).

We conclude that at pressures below 55 GPa and temperatures below 2,000 K, corundum and iron do not react. At higher pressures (above 65 GPa) and temperatures (above 2,000 K) we observed a chemical reaction which can be described in general as



where x is varied from 0.02–0.03 to 0.25.

In other words, at the conditions of the Earth's lower mantle, iron can reduce aluminium oxide to metallic aluminium, which eventually forms an alloy with the rest of iron.

Pressure is known to play an important part in the direction of oxidation–reduction reactions, particularly in the Fe–O system²⁰. Our study reveals an extreme case: iron, which at ambient

conditions is more electronegative than aluminium, changes its chemical nature at high pressures and temperatures and is able to reduce aluminium from its oxide. It also means that at the core–mantle boundary iron can reduce not only silicon^{6,9}, but also aluminium from aluminosilicates. Our results show that complex chemical reactions will occur in the D'' layer forming chemical heterogeneities composed of regions enriched in iron, or in silicon and/or aluminium. □

Received 24 November 2000; accepted 21 June 2001.

- Vidale, J. E. & Earle, P. S. Fine-scale heterogeneity in the Earth's inner core. *Nature* **404**, 273–275 (2000).
- Vidale, J. E., Dodge, D. A. & Earle, P. S. Slow differential rotation of the Earth's inner core indicated by temporal changes in scattering. *Nature* **405**, 445–448 (2000).
- Breger, L., Romanowicz, B. & Rousset, S. New constraints on the structure of the inner core from P'P'. *Geophys. Res. Lett.* **27**, 2781–2784 (2000).
- Carnero, E. J. & Jeanloz, R. Fuzzy patches on the Earth's core–mantle boundary? *Geophys. Res. Lett.* **27**, 2777–2780 (2000).
- Stixrude, L. & Brown, J. M. The Earth's core. *Rev. Mineral.* **37**, 262–283 (1998).
- Knittle, E. & Jeanloz, R. Simulating the core–mantle boundary: an experimental study of high-pressure reactions between silicates and liquid iron. *Geophys. Res. Lett.* **16**, 609–612 (1989).
- Poirier, J. P., Malavergne, V. & Le Mouél, J. L. in *The Core–Mantle Boundary Region* (eds Gurnis, M., Wyssession, M. E., Knittle, E. & Buffet, B. A.) 131–137 (AGU, Washington, 1998).
- Knittle, E. & Jeanloz, R. Earth's core–mantle boundary: Results of experiments at high pressures and high temperatures. *Science* **251**, 1438–1443 (1991).
- Goarant, F., Guyot, F., Peyroneau, J. & Poirier, J. P. High-pressure and high-temperature reactions between silicates and liquid iron alloys in the diamond anvil cell, studied by analytical electron microscopy. *J. Geophys. Res.* **97**, 4477–4487 (1992).
- Song, X. & Ahrens, T. J. Pressure–temperature range of reactions between liquid iron in the outer core and mantle silicates. *Geophys. Res. Lett.* **21**, 153–156 (1994).
- Malavergne, V. *Etude expérimentale à haute pression et haute température du partage des éléments de transition entre les phases du manteau et du noyau de la Terre*. Thesis, Univ. Paris 7, (1996).
- Tschauner, O. *et al.* Partitioning of nickel and cobalt between silicate perovskite and metal at pressures up to 80 GPa. *Nature* **398**, 604–607 (1999).
- Anderson, D. L. *Theory of the Earth* (Blackwell Scientific, Boston, 1989).
- Allègre, C. J., Poirier, J.-P., Humler, E. & Hofmann, A. W. The chemical composition of the Earth. *Earth Planet. Sci. Lett.* **134**, 515–526 (1995).
- Ruff, L. J. & Anderson, D. L. Core formation, evolution, and convection; a geophysical model. *Phys. Earth Planet. Inter.* **21**, 181–201 (1980).
- McCammon, C. A. Perovskite as a possible sink for ferric iron in the lower mantle. *Nature* **387**, 694–696 (1997).
- Lauterbach, S., McCammon, C. A., van Aken, P., Langenhorst, F. & Seifert, F. Mössbauer and ELNES spectroscopy of (Mg,Fe)(Si,Al)O₃ perovskite: a highly oxidised component of the lower mantle. *Contrib. Mineral. Petrol.* **138**, 17–26 (2000).
- Hilgert, V. J., Drake, M. J. & Rubie, D. C. High-pressure and high-temperature experiments on core–mantle segregation in the accreting Earth. *Science* **264**, 1441–1445 (1994).
- Li, J. & Agee, C. B. Geochemistry of mantle–core differentiation at high pressure. *Nature* **381**, 686–689 (1996).
- Fabrichnaya, O. B. & Sundman, B. The assessment of thermodynamic parameters in the Fe–O and Fe–Si–O systems. *Cosmochim. Acta* **61**, 4539–4555 (1997).
- Dubrovinsky, L. S., Saxena, S. K., Lazor, P. & Weber, H.-P. Structure of β -iron at high temperature and pressure. *Science* **281**, 11–12 (1998).
- Yoo, C. S., Akella, J., Campbell, A. J., Mao, H. K. & Hemley, R. J. Phase diagram of iron by in situ X-ray diffraction: implications for Earth's core. *Science* **270**, 1473–1475 (1995).
- El Goresy, A., Dubrovinsky, L., Sharp, T. G. & Chen, M. A new monoclinic post-stishovite polymorph of silica in the SNC meteorite Shergotty. *Science* **288**, 1632–1634 (2000).
- Dubrovinsky, L. S. *et al.* Stability of ferropervicite in the lower mantle. *Science* **289**, 430–432 (2000).
- Dubrovinsky, L. S. & Saxena, S. K. Emissivity measurements on same metals and oxides using multiwavelength spectral radiometry. *High Pressure High Temp.* **31**, 393–399 (1999).
- Mao, H. K., Bell, P. M. & Hadidiacos, C. in *High-Pressure Research in Mineral Physics* (eds Manghnani, M. H. & Syono, Y.) 135–138 (Terra Scientific, Tokyo, and American Geophysical Union, Washington DC, 1987).
- Dubrovinsky, L. S., Saxena, S. K. & Lazor, P. High-pressure and high-temperature in situ x-ray diffraction study of iron and corundum to 68 GPa using internally heated diamond anvil cell. *Phys. Chem. Minerals* **25**, 434–441 (1998).
- Dubrovinskaya, N. A., Dubrovinsky, L. S., Karlsson, A., Saxena, S. K. & Sundman, B. Experimental study of thermal expansion and phase transformations in iron-rich Fe–Al alloys. *Calphad* **23**, 69–84 (1999).
- Pasternak, M. P. *et al.* High pressure collapse of magnetism in Fe_{0.94}O: Mössbauer spectroscopy beyond 100 GPa. *Phys. Rev. Lett.* **79**, 5046–5049 (1997).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London office of Nature.

Acknowledgements

We thank D. C. Rubie for comments and discussions. This work was supported by the Swedish Science Foundation (NFR), the Wallenberg and Göran Gustafsons Stiftelse Funds.

Correspondence and requests for materials should be addressed to L.D. (e-mail: Leonid.Dubrovinsky@geo.uu.se).

The last of the dinosaur titans: a new sauropod from Madagascar

Kristina Curry Rogers* & Catherine A. Forster

Department of Anatomical Sciences, Health Sciences Center, State University of New York, Stony Brook, New York 11794, USA

The Titanosauria, the last surviving group of the giant sauropod dinosaurs, attained a near-global distribution by the close of the Cretaceous period (65 Myr ago). With the exception of a few new discoveries in Argentina^{1–3}, most titanosaurs are known only from fragmentary postcranial skeletons and rare, isolated skull elements^{4–9}. Here we describe the most complete titanosaur yet discovered. *Rapetosaurus krausei* gen. et sp. nov., from the Maevarano Formation of Madagascar, provides a view of titanosaur anatomy from head to tail. A total-evidence phylogenetic analysis supports a close relationship between brachiosaurids and titanosaurs (Titanosauriformes^{10–13}). The inclusion of cranial data from *Rapetosaurus* also lays to rest questions concerning the phylogeny of the enigmatic Mongolian genera *Nemegtosaurus* and *Quaesitosaurus*^{14,15}. In spite of their elongated, diplodocoid-like skulls, all three taxa are now firmly nested within Titanosauria.

Fragmentary titanosaur remains have been known from the Upper Cretaceous strata of Madagascar since Charles Depéret described '*Titanosaurus madagascariensis*' from two procoelous caudal vertebrae, a partial humeral diaphysis and a large osteoderm¹⁶. Currently we recognize two titanosaur taxa in the Maevarano Formation that are distinguished, in part, by their caudal vertebrae^{17–19}. Depéret's *T. madagascariensis* syntype includes both caudal morphologies and is therefore considered a *nomen dubium*.

The disarticulated bones of the holotype skull of *Rapetosaurus krausei* were recovered from a single stratum over an area of ~1 m² at locality MAD 96-02. Holotype material shows no duplication of elements and exhibits exact sutural articulations. Additional skull material from at least two juvenile sauropods was recovered from locality MAD 93-18 (refs 18, 19). Most of these juvenile skull elements were found intimately associated with a nearly complete juvenile postcranial skeleton that also shows no duplication of elements. Several of the juvenile skull elements from MAD 93-18 share autapomorphies with the holotype adult skull, and thus are referred to *R. krausei*.

Dinosauria Owen, 1842

Saurischia Seeley, 1888

Sauropoda Marsh, 1878

Titanosauria Bonaparte & Coria, 1993

Rapetosaurus krausei gen. et sp. nov.

Etymology. *Rapeto* (ruh-PAY-tu, Malagasy), a mischievous giant of Malagasy folklore; *saurus* (Greek), lizard. Specific name for David W. Krause, in recognition of his contributions to palaeontology in Madagascar.

Holotype. Université d'Antananarivo UA 8698, adult skull including right maxilla with eight teeth, left maxilla, right lacrimal, left jugal, right and left nasals, right quadrate, right and left pterygoids, partial basioccipital, right paroccipital process, left dentary with 11 teeth, both angulars, right surangular, and five associated teeth (Figs 1 and 2).

Referred specimens. Field Museum of Natural History FMNH PR 2184–2192, 2194, 2196, 2197, 2209, 2210; right exoccipital, opisthotic, laterosphenoid, supraoccipital, associated right and left frontals,

two right prefrontals, left surangular, right parietal, left squamosal, right quadrate, right pterygoid, right angular, six associated teeth, fused basioccipital, basisphenoid, parasphenoid, associated juvenile skeleton, mid-caudal centrum, respectively; UCB (Université Claude Bernard) 92829, mid-caudal centrum (Figs 1–3).

Type locality and horizon. MAD 96-02, Mahajanga basin, north-western Madagascar; Anembalemba Member, Maevarano Formation, Upper Cretaceous, Maastrichtian²⁰.

Diagnosis. *Rapetosaurus* is characterized by the following autapomorphies: expanded antorbital fenestra extends over tooth row; preantorbital fenestra positioned posterior to antorbital fenestra; subnarial foramen anteriorly positioned and dorsoventrally elongate; jugal process of maxilla posterodorsally elongate and narrow; frontals with median dome; quadrate with V-shaped quadratojugal articulation; supraoccipital with two anteriorly directed median parietal processes; pterygoid with extremely shallow basiptyergoid articulation and dorsoventrally expanded anterior process; basiptyergoid processes diverge only at distal extremes; dentary with 11 alveoli that extend two-thirds the length of the element; gracile cylindrical teeth with high-angle planar wear facets; 16 cervical vertebrae with constricted neural canals and continuous pre- and postspinal coels devoid of pre- or postspinal laminae; cervical neural spines with proximal bifurcation and three pneumatized coels bounded by discrete laminae; 11 dorsal vertebrae with deep lateral pleurocoels; dorsal neural spines with strong pre- and postspinal laminae in deeply excavated anterior and posterior coels; dorsals with median interpre- and interpostzygapophyseal laminae; middle and posterior dorsals with divided spinodiapophyseal lamina; six sacral centra with deep lateral pleurocoels; all caudal centra procoelous with convex ventral margin lacking

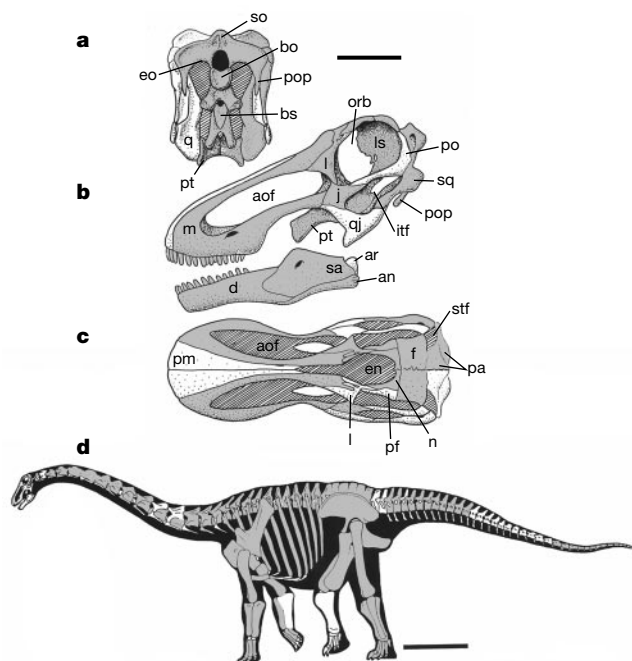


Figure 1 *Rapetosaurus krausei* from the Upper Cretaceous Maevarano Formation of Madagascar. The skull reconstruction is a composite based on holotype and referred specimens and shown in posteroventral (a), left lateral (b) and dorsal (c) views. The skeleton, in left lateral view (d), is based on a juvenile postcranial skeleton (FMNH PR 2209), which is ~75% complete. Scale bars are 10 cm (a–c) and 1 m (d). an, angular; aof, antorbital fenestra; ar, articular; bo, basioccipital; bs, basisphenoid; d, dentary; en, external nares; eo, exoccipital; f, frontal; itf, infratemporal fenestra; j, jugal; l, lacrimal; ls, laterosphenoid/orbitosphenoid; m, maxilla; n, nasal; orb, orbit; pa, parietal; pop, paroccipital process; pf, prefrontal; pm, premaxilla; po, postorbital; pt, pterygoid; q, quadrate; qj, quadratojugal; sa, surangular; so, supraoccipital; sq, squamosal; stf, supratemporal fenestra.

* Present address: Department of Paleontology, Science Museum of Minnesota, 120 W. Kellogg Boulevard, St Paul, Minnesota 55102, USA; and Macalester College, Geology Department, St Paul, Minnesota 55105, USA.

excavation; anterior caudal centra broad transversely and antero-posteriorly compressed; middle-posterior caudal centra with constant length:width ratio; anterior-middle caudal neural spines with spinoprezygapophyseal, prespinal and postspinal laminae on rectangular and anteriorly positioned neural arches; chevrons throughout 80% of tail; iliac peduncle of ischium comprises one-quarter of acetabulum; ischial peduncle of ilium low and poorly developed; pubis more than twice as long as ischium; scapula and coracoid with equal glenoid contribution; scapular blade not distally expanded; humerus/femur length quotient 0.80; radius and ulna with oblique interosseus ridges (Figs 1–3).

The *Rapetosaurus* maxilla has a posterodorsally sloping dorsal ascending process with a long premaxillary articulation along its proximal two-thirds. The enlarged antorbital fenestra is unique

among sauropods in that it extends anteriorly over the tooth row and is bound ventrally by an elongate, narrow jugal process (Figs 1a–c and 2a). Cylindrical teeth occupy the anterior two-thirds of the element, and have high-angled, lingually oriented wear facets (Fig. 2b) indicating precise tooth-to-tooth occlusion. This straight maxillary profile and the cylindrical teeth are more similar to diplodocoid skulls than they are to the ‘stepped’ snout profile of brachiosaurids. In contrast to diplodocoids, *Rapetosaurus* teeth occupy the anterior two-thirds of the maxilla. The external nares of *Rapetosaurus* are retracted to the level of the orbit and incompletely divided by medial nasal processes (Figs 1c and 2e) as in diplodocoids, thus confirming that the nostrils were located on the top of the skull. As in other titanosaurs, the supratemporal fenestrae are transversely oriented, and the quadrate is hollow, with a posteriorly facing excavation (Fig. 2f). *Rapetosaurus* has short basal tubera that diverge at 35° and project away from the skull. The basiptyergoid processes are longer than those of any other titanosaur, and fit into shallow facets on the plate-like pterygoids, as in *Nemegtosaurus*.

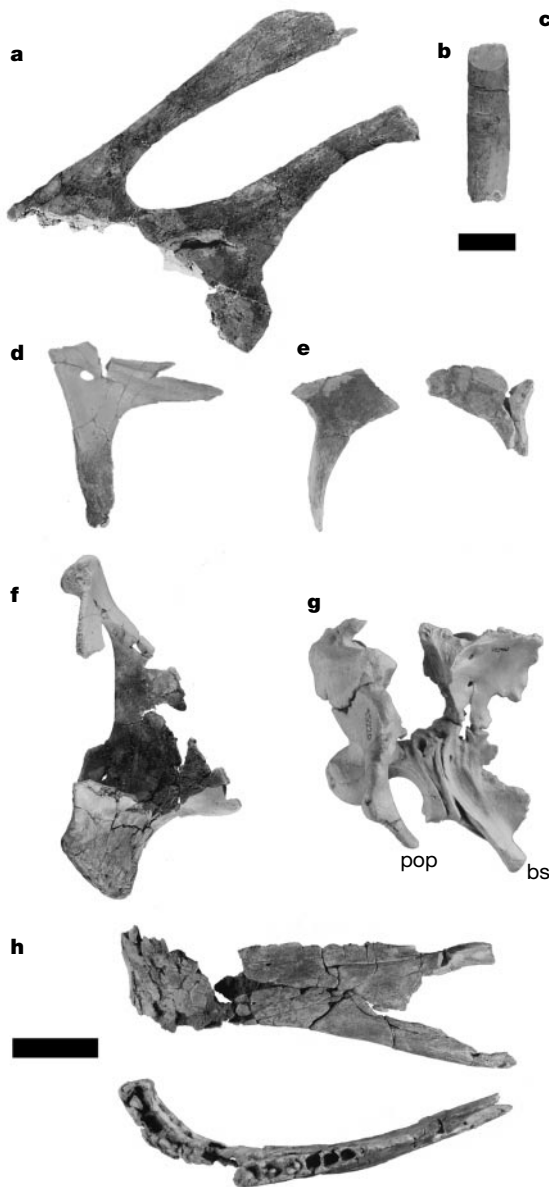


Figure 2 *Rapetosaurus krausei* cranial elements. **a**, Holotype (UA 8698) left maxilla in lateral view. **b**, UA 8698 worn maxillary tooth in lingual view showing high-angled wear facet. **c**, UA 8698 unworn maxillary tooth in mesial/distal view. **d**, UA 8698 right lacrimal in lateral view. **e**, UA 8698 right and left nasals in dorsal view. **f**, UA 8698 right quadrate in anterior view. **g**, FMNH-PR 2184, 2197, articulated supraoccipital, exoccipital/opisthotic, basioccipital, basiptyergoid, basisphenoid and laterosphenoid in right lateral view. **h**, UA 8698 dentary in left lateral (top) and dorsal (bottom) views. Scale bars are 5 cm (**a**, **d**–**h**) and 1 cm (**b**, **c**). bs, basisphenoid; pop, paraoccipital process.

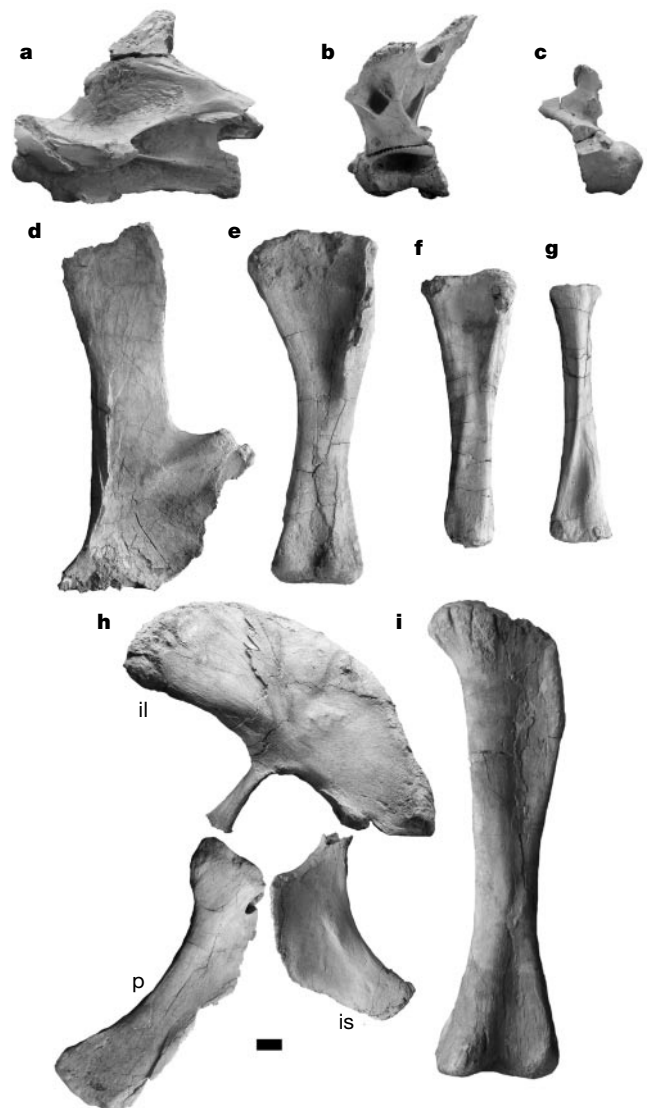


Figure 3 *Rapetosaurus krausei* postcranial elements, all from referred juvenile skeleton (FMNH PR 2209). **a**, Cervical vertebra 10 in left lateral view. **b**, Dorsal vertebra 4 in left lateral view. **c**, Anterior caudal vertebra in left lateral view. **d**, Left scapula in lateral view. **e**, Left humerus in anterior view. **f**, Left ulna in anterior view. **g**, Left radius in posterior view. **h**, Left pubis, ischium and ilium in lateral view. **i**, Left femur in anterior view. Scale bar is 3 cm. il, ilium; is, ischium; p, pubis.

Antarctosaurus, *Saltasaurus*, *Malawisaurus* and *Rapetosaurus* share elongate, recurved paroccipital processes (Figs 1a, b and 2g). The lower jaw is similar to that of brachiosaurids in its broad U shape, the 11 alveoli occupying the anterior two-thirds of the dentary, and the surangular's prominent coronoid eminence (Figs 1c and 2h). Based on the associated elements of the lower jaw, the skull of an adult *Rapetosaurus* is ~40 cm long.

As in other titanosaurs, the cervical vertebrae (Fig. 3a) are elongate with shallow lateral pleurocoels. The cervical neural canal is narrow and constricted and the proximal neural spines are bifid. Cervical neural spines and pleurocoels are marked by large, polished areas indicating axial pneumaticity. Like *Alamosaurus* and *Titanosaurus colberti*, dorsal neural spines (Fig. 3b) are craniocaudally compressed and broad, tapering only at their distal extremes. Posterior dorsal transverse processes are positioned directly above the parapophysis. In contrast to most other titanosaurs, *Rapetosaurus* dorsals and sacrales have deep pleurocoels. All caudal vertebrae (Fig. 3c) are strongly procoelous, as in *T. colberti* and *Saltasaurus*. Unique among titanosaurs, *Rapetosaurus* bears a strongly developed spinoprezygapophyseal lamina on anterior caudal neural spines.

The appendicular skeleton of *Rapetosaurus* is characterized by many typical titanosaur features (Fig. 3d–i), including a lateral projection from the proximal end of the femur, elongated metacarpals, an extremely elongated pubis, an anterodorsally expanded and semicircular preacetabular process of the ilium, and an elevated olecranon process of the ulna. Like most saltasaurines, the scapular blade of *Rapetosaurus* forms a 45° angle with the coracoid. The distal condyles of the humerus and femur expand onto the anterior

surface of the shaft in *Rapetosaurus*, *Neuquensaurus* and *Saltasaurus*.

Discussion. *Rapetosaurus krausei* exhibits a suite of features (Fig. 4) that confirm its status as a titanosaur^{10–13}. Titanosaurs have historically been allied with diplodocoid sauropods^{21–25}, but recent studies suggest they share a closer relationship with brachiosaurids within Titanosauriformes^{10–13}. Character support for Titanosauriformes has rested almost exclusively on postcranial data, primarily owing to the lack of titanosaur skulls. *Rapetosaurus* offers an opportunity to test the monophyly of Titanosauriformes with abundant, associated skull and skeletal data. *Rapetosaurus* also offers the best opportunity yet to resolve the phylogenetic position of two controversial Mongolian sauropods, *Nemegtosaurus* and *Quaesitosaurus*. These taxa are known only from skulls, and have been alternately placed within the Titanosauria^{11,12} or the Diplodocoidea^{5,13,15,24}. With regard to these Mongolian genera, *Rapetosaurus* is particularly illuminating because its elongated skull with retracted nares exhibits a general similarity to diplodocoids, whereas its skeleton shows striking commonalities with brachiosaurids and other titanosaurs.

We conducted a phylogenetic analysis with 228 characters (74 cranial and 154 postcranial; see Supplementary Information) and 16 in-group taxa (Fig. 4)^{10–13,15,26}. Our analysis supports the higher-level grouping of Titanosauriformes (the most recent common ancestor of *Brachiosaurus* and *Saltasaurus*, and all of its descendants^{10,12}), which are united by eleven characters including four unambiguous characters: 23, presence of a rod-like palatine–maxillary contact; 106, simple and undivided chevron blades; 152, proximal one-third of femur deflected medially; 219, blades of middle and caudal chevrons curve backward and downwards.

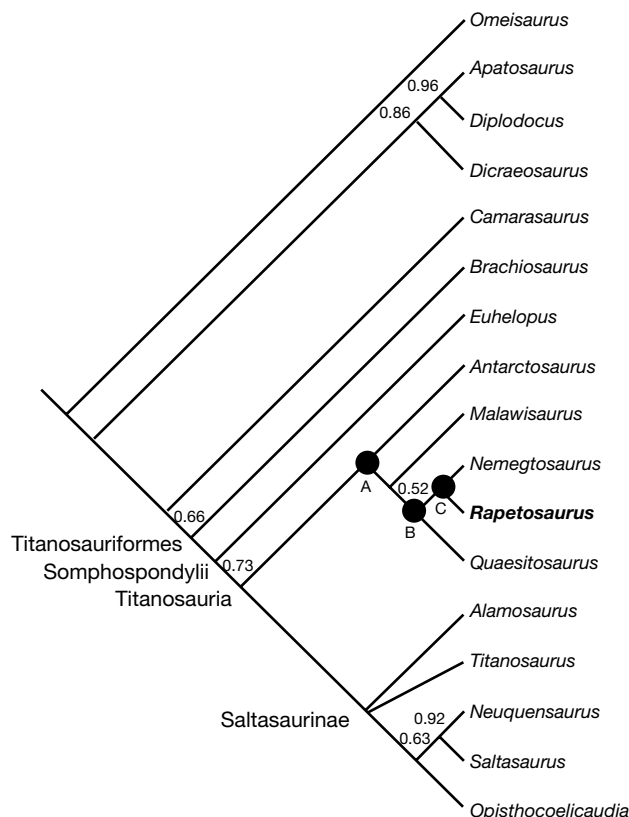


Figure 4 Cladogram showing phylogenetic position of *Rapetosaurus krausei* based on strict consensus of the two most parsimonious trees (447 steps; consistency index = 0.5481; retention index = 0.6189) generated by a branch and bound search in PAUP* (v.4.0b2a)³¹. Trees were rooted with *Omeisaurus* as the outgroup and optimized with delayed transformations, and all characters were unordered. Characters were drawn from recent analyses of sauropod phylogeny^{10–13,15,24,26}. *Titanosaurus* skull data is

derived from *T. indicus*, postcranial data from *T. colberti*; *Antarctosaurus* refers to *A. septentrionalis*. Bootstrap values (300 replicates) are listed at nodes. Bremer support values for labelled nodes are as follows: Nodes A–C, 1; Titanosauria, 2; Titanosauriformes, 3. For character list and taxon/character matrix, see Supplementary Information.

Sixteen additional steps are required to move Titanosauria into its more traditional position as a sister group to diplodocoids. Titanosauria is united by 20 characters, including nine that are unambiguous: 71, hyposphene–hypantra articulations present on posterior dorsal neural arches; 77, spongy bone texture of caudals; 78, ≤ 35 caudal vertebrae; 108, haemal canal approximately half of chevron length; 118, crescentic sternal plates; 125, ulnar olecranon process prominent and projects above the proximal articulation; 128, distal breadth of radius twice midshaft breadth; 154, transverse axis of femoral distal condyles bevelled dorsomedially $\sim 10^\circ$; 226, ischium/pubis quotient ≤ 0.90 . Within Titanosauriformes, two clades of titanosaurs are distinguished: a group that includes *Rapetosaurus* (*Rapetosaurus* clade) and the Saltasaurinae²⁷ (Fig. 4). The Saltasaurinae is united by 16 characters including four that are unambiguous: 80, first caudal centrum with biconvex articular shape; 121, deltopectoral crest reduced to a low rounded crest or ridge; 138, manual digits II and III lack phalanges; 157, distal breadth of tibia more than twice its midshaft breadth.

Taxa in the *Rapetosaurus* clade are united by four ambiguous cranial synapomorphies (Fig. 4, Node A): 12, frontals with fused midline contact; 13, parietal does not contribute to post-temporal fenestra; 16, occipital region of skull anteroposteriorly deep with paroccipital processes oriented posterolaterally (reversal); 25, basiptyergoid process at least four times that of basal diameter. All four of these characters are shared with one or more diplodocoids, and the addition of only one step collapses the node. Within the *Rapetosaurus* clade, *Malawisaurus* is the sister taxon of *Quaesitosaurus* (*Rapetosaurus* + *Nemegtosaurus*), a relationship supported by 10 ambiguous characters, including four postcranial features shared with one or more diplodocoids. *Nemegtosaurus*, *Quaesitosaurus*, and *Rapetosaurus* are further united by eleven synapomorphies, two of which are unambiguous (Fig. 4, Node B): 37, posterodorsal process of the splenial present; 195, angle between the long axis of mandibular symphysis and the long axis of mandible $\sim 90^\circ$. *Rapetosaurus* and *Nemegtosaurus* are united by seven ambiguous cranial characters, five of which occur in one or more diplodocoids (Fig. 4, Node C): 5, external nares are retracted to a position between the eyes; 9, presence of an anterior prefrontal process; 11, frontal contributes to supratemporal fenestra; 172, external nares face dorsally or rostradorsally; 174, absence of a 'stepped' snout profile; 190, ectopterygoid process of pterygoid lies ventral or caudoventral to lacrimal; 197, slenderness indices for teeth > 5.0 .

All five taxa in the *Rapetosaurus* clade, but only two saltasaurines (*Saltasaurus* and *Titanosaurus*), have cranial material. When only postcranial data are analysed (*Nemegtosaurus*, *Quaesitosaurus* and *Antarctosaurus* are eliminated owing to their lack of postcranial data), *Rapetosaurus* + *Malawisaurus* remains the sister taxon to Saltasaurinae. Whereas the current data seem to indicate a close relationship between *Rapetosaurus* and the two Mongolian taxa, missing data and the uneven distribution of cranial and postcranial materials among taxa lead us to advise caution when interpreting such relationships. For example, the two unambiguous characters that unite *Rapetosaurus*, *Nemegtosaurus* and *Quaesitosaurus* (Node B; characters 37 and 195) are coded as missing for all saltasaurines and *Antarctosaurus*. Even so, a rearrangement of *Rapetosaurus*, *Nemegtosaurus* and *Quaesitosaurus* to a sister-group relationship with diplodocoids results in the addition of 29 steps.

From a palaeobiogeographic perspective, a resolved phylogeny of the globally distributed Titanosauria offers an opportunity to explore Cretaceous vicariance. Our phylogenetic analysis suggests that the biogeographic history of *Rapetosaurus* may be most closely linked to Africa, Asia and India. This finding contrasts with the proposed biogeographic history of Maevarano Formation mammals, crocodyliforms and theropod dinosaurs, which make a strong case for extensive faunal interchange between South America, India and Madagascar²⁸. However, missing data coupled with low Bremer

support and bootstrap values for most nodes within Titanosauria, as well as the lack of a palaeogeographic model to explain the potential Asia–Madagascar connection, reduces our confidence in this arrangement. As such, the significance of these biogeographic possibilities must be tested in the context of a more global analysis of titanosaur phylogeny and, we hope, with the addition of more cranial and postcranial data.

Finally, titanosaurs are unique among sauropod clades in that they evolved alongside large-bodied, herbivorous ornithischian dinosaurs, notably the horned ceratopsians, 'duck-billed' hadrosaurs and armoured ankylosaurs^{27,29}. *Rapetosaurus*, in contrast, shares the Mahajanga basin with only one large-bodied herbivore, another titanosaur^{17–19}. Small herbivores are also exceedingly rare in the Cretaceous of Madagascar, with only a single herbivorous/omnivorous crocodyliform³⁰ discovered in more than 100 years of collection. The occurrence of these co-existing titanosaurs in the apparent absence of ornithischian dinosaurs suggests a different herbivore community dynamic on Madagascar than is seen elsewhere in the Cretaceous. □

Received 4 March; accepted 25 May 2001.

- Calvo, J. O., Coria, R. A. & Salgado, L. Uno de los más completos titanosáuridos (Dinosauria—Sauropoda) registrados en el mundo. *Ameghiniana* **34**, 534 (1997).
- Coria, R. A. & Salgado, L. Nuevos aportes a la anatomía craneana de los saurópodos titanosáuridos. *Ameghiniana* **36**, 98 (1999).
- Martínez, R. An articulated skull and neck of Sauropoda (Dinosauria: Saurischia) from the Upper Cretaceous of Central Patagonia, Argentina. *J. Vert. Paleontol.* **18**, 61A (1998).
- Berman, D. S. & Jain, S. L. The braincase of a small sauropod dinosaur (Reptilia: Saurischia) from the Upper Cretaceous Lameta Group, Central India, with a review of Lameta Group localities. *Ann. Carnegie Mus.* **51**(21), 405–422 (1982).
- Nowinski, A. *Nemegtosaurus mongoliensis* n. gen. n. sp. (Sauropoda) from the uppermost Cretaceous of Mongolia. *Palaentol. Pol.* **25**, 57–81 (1971).
- Kurzanov, S. M. & Bannikov, A. F. A new sauropod from the Upper Cretaceous of Mongolia. *Paleontol. J.* **2**, 90–96 (1983).
- Powell, J. E. in *Los Dinosaurios y su entorno biótico. Actas 2º Curso de Paleontología en Cuenca* (eds Sanz, J. L. & Buscalioni, A. S.) 165–230 (Instituto 'Juan de Valdes', Cuenca, 1992).
- Chatterjee, S. & Rudra, D. K. KT events in India: impact, rifting, volcanism and dinosaur extinction. *Mem. Queensland Mus.* **39**, 489–532 (1996).
- Gomani, E. Sauropod cranial elements from Malawi: implications for titanosaurid cranial morphology. *J. Vert. Paleontol.* **18**, 46A (1998).
- Salgado, L., Coria, R. A. & Calvo, J. O. Evolution of titanosaurid sauropods. I: Phylogenetic analysis based on the postcranial evidence. *Ameghiniana* **34**, 3–32 (1997).
- Salgado, L. & Calvo, J. O. Evolution of titanosaurid sauropods. II: The cranial evidence. *Ameghiniana* **34**, 33–48 (1997).
- Wilson, J. A. & Sereno, P. C. Early evolution and higher-level phylogeny of sauropod dinosaurs. *J. Vert. Paleontol.* (Suppl.) **18**, 1–68 (1998).
- Upchurch, P. The phylogenetic relationships of sauropod dinosaurs. *Zool. J. Linn. Soc.* **124**, 43–103 (1998).
- Calvo, J. O., Salgado, L. & Coria, R. A. Nueva reconstrucción del cráneo de *Nemegtosaurus mongoliensis*, Nowinski, 1971 (Sauropoda). *Acta Geol. Lilloana* **13**, 156–157 (1998).
- Upchurch, P. The phylogenetic relationships of the Nemegtosauridae (Saurischia: Sauropoda). *J. Vert. Paleontol.* **19**, 106–125 (1999).
- Depéret, C. Note sur les Dinosauriens Sauropodes et Théropodes dans le Crétacé supérieur de Madagascar. *Bull. Soc. Géol. Fr.* **21**, 176–194 (1896).
- Ravoavy, F. Identification et mise en catalogue des vertèbres fossiles récoltes dans le Crétacé supérieur continental de la région de Berivotra (Majunga) fouille 1987. *Univ. d'Antananarivo Mém. Rech.* **II**, 55–104 (1991).
- Curry Rogers, K. & Forster, C. A. in *VII Int. Symp. Mesozoic Terrestrial Ecosystems* 20–21 (Museo de Ciencias Naturales, Buenos Aires, 1999).
- Curry Rogers, K. & Forster, C. A. New sauropods from Madagascar: a glimpse into titanosaur cranial morphology and evolution. *J. Vert. Paleontol.* **19**, 41A (1999).
- Rogers, R. R., Hartman, J. H. & Krause, D. W. Stratigraphic analysis of Upper Cretaceous rocks in the Mahajanga Basin, northwestern Madagascar: implications for ancient and modern faunas. *J. Geol.* **108**, 275–301 (2000).
- Huene, F. von Die Besonderheit der Titanosaurier. *Centralblatt Mineral. Geol. Palaontol.* **B**, 493–499 (1929).
- Romer, A. S. *Osteology of the Reptiles* (Univ. Chicago Press, Chicago, 1956).
- Gauthier, J. in *The Origin of Birds and the Evolution of Flight* (ed. Padian, K.) 1–55 (California Academy of Sciences, San Francisco, 1986).
- Upchurch, P. Evolutionary history of sauropod dinosaurs. *Phil. Trans. R. Soc. Lond. B* **349**, 365–390 (1995).
- McIntosh, J. S. in *The Dinosauria* (eds Weishampel, D. B., Dodson, P. & Osmólska, H.) 345–401 (Univ. California Press, Berkeley, 1990).
- Wilson, J. A. *The Evolutionary History of the Sauropoda*. Thesis, Univ. Chicago (1999).
- Powell, J. E. *Revisión de los Titanosáuridos de América del Sur*. Thesis, Univ. Nacional Tucumán (1986).
- Krause, D. W. & et al. The Late Cretaceous vertebrate fauna of Madagascar: implications for Gondwanan paleobiogeography. *GSA Today* **9**(8), 1–7 (1999).
- Coria, R. A. Ornithomimid dinosaurs from the Neuquén Group, Patagonia, Argentina: phylogeny and biostratigraphy. *Natl Sci. Mus. Monogr.* **15**, 47–60 (1999).

30. Buckley, G. B., Brochu, C. B., Krause, D. W. & Pol, D. A pug-nosed crocodyliiform from the Late Cretaceous of Madagascar. *Nature* **405**, 941–944 (2000).
31. Swofford, D. L. *PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods)* Version 4.0b2a. (Sinauer, Sunderland, 1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

'Misaotra betsika' to the villagers of Berivotra, Madagascar. M. Hallet provided the illustrations and M. Stewart and T. Ready took the photographs. Thanks to R. Rogers, D. Krause, M. O'Leary, R. Coria and P. Upchurch for comments and discussion. Funding was provided by NSF, National Geographic Society, Jurassic Foundation, GSA and the Gabor Inke Graduate Student Fellowship.

Correspondence and requests for materials should be addressed to K.C.R. (e-mail: krogers@smm.org).

Neanderthal cranial ontogeny and its implications for late hominid diversity

Marcia S. Ponce de León & Christoph P. E. Zollikofer

Anthropologisches Institut and Institut für Informatik/MultiMedia Laboratorium Universität Zürich-Irchel, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland
The authors contributed equally to this work

Homo neanderthalensis has a unique combination of craniofacial features that are distinct from fossil and extant 'anatomically modern' *Homo sapiens* (modern humans). Morphological evidence, direct isotopic dates¹ and fossil mitochondrial DNA from three Neanderthals^{2,3} indicate that the Neanderthals were a separate evolutionary lineage for at least 500,000 yr. However, it is unknown when and how Neanderthal craniofacial autapomorphies (unique, derived characters) emerged during ontogeny. Here we use computerized fossil reconstruction⁴ and geometric morphometrics^{5,6} to show that characteristic differences in cranial and mandibular shape between Neanderthals and modern humans arose very early during development, possibly prenatally, and were maintained throughout postnatal ontogeny. Postnatal differences in cranial ontogeny between the two taxa are characterized primarily by heterochronic modifications of a common spatial pattern of development. Evidence for early ontogenetic divergence together with evolutionary stasis of taxon-specific patterns of ontogeny is consistent with separation of Neanderthals and modern humans at the species level.

Comparative analyses of immature crania indicate that diagnostic Neanderthal characters appeared early during ontogeny^{7,8} and that the Neanderthal ontogenetic process was fast relative to that of the modern humans^{7,9–11}. Here, we use a new methodological approach to study the comparative ontogeny of Neanderthal and modern human skulls. After computerized reconstruction of fragmentary fossil specimens^{4,12–14}, we applied geometric morphometric methods (GMM)⁵ to identify and visualize complex patterns of morphological change during ontogeny (see Methods). In GMM the form of a specimen is described by the spatial configuration of a set of three-dimensional anatomical landmarks. Size-corrected variation in shape can then be computed in terms of between-specimen rearrangements of landmark positions. To capture large trends in shape variation in ontogenetic samples of Neanderthals and modern humans, we used relative warp analysis^{5,6}, which separates shape variability into statistically independent factors. Each relative warp thus captures an independent aspect of shape variation in the sample (Fig. 1) that can be plotted ontogenetically

as temporal (Fig. 2) and spatial (Figs 3 and 4) patterns of morphological change.

Our relative warp analyses are based on a large cross-sectional ontogenetic series of Neanderthal cranial and mandibular specimens that were reconstructed by computerized methods, and a comparative fossil/recent modern human sample. The Neanderthal sample mostly comprises individuals from dental stage 3 (3–6 yr) to adulthood, but includes an early postnatal mandible (Amud 7, about 0.5 yr⁸) and a cranium and mandible from dental stage 2 (Pech de l'Azé, about 2.5 yr); the modern human sample includes individuals from all ontogenetic stages—from perinatal through to adulthood (see Methods and Supplementary Information for sample details and landmark definitions). We computed the statistically independent relative warps for the combined craniomandibular landmark configurations (Fig. 1a) and for cranial and mandibular configurations in isolation (Fig. 1b, c), and plotted these against three additional factors: (1) individual age (dental

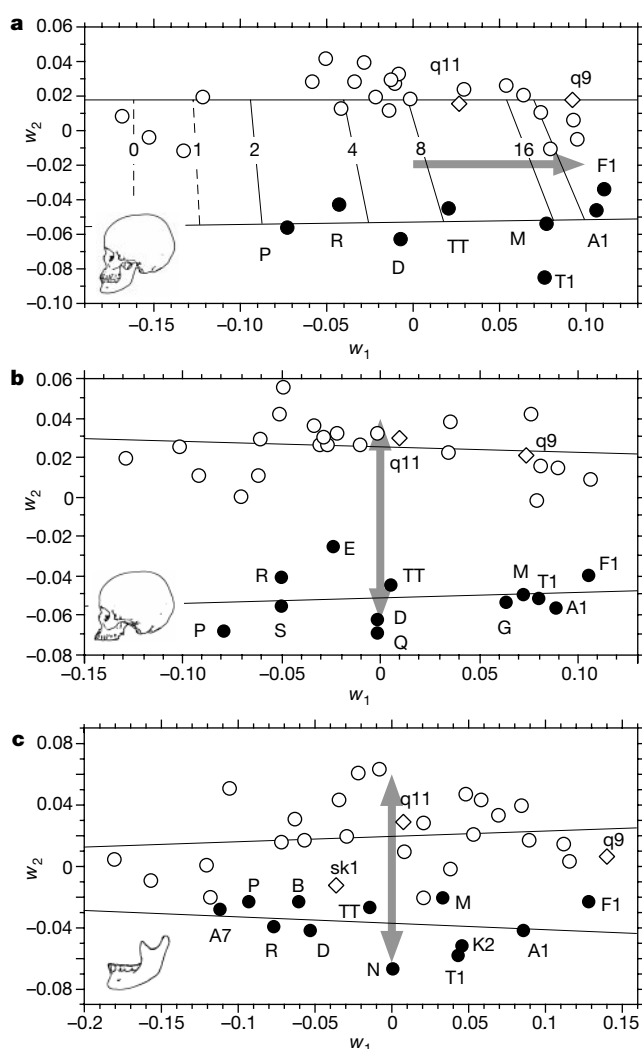


Figure 1 Shape variability in an ontogenetic series of Neanderthals (filled circles; see Methods for specimen labels) and modern humans (open circles/diamonds indicate extant/fossil specimens, respectively) for craniomandibular (a), cranial (b) and mandibular (c) landmark configurations. The labels w_1 and w_2 represent factors of shape variation resulting from relative warp analysis^{5,6}. Neanderthals and modern humans follow ontogenetic trajectories that are approximately parallel (lines are principal axes of within-taxon correlation). The arrow in a characterizes the shared Neanderthal/modern human mode of development as shown in Fig. 3; taxon-specific trajectories are connected with lines at dental ages 0 (birth), 1, 2, 4, 8, 16 yr and adulthood. Arrows in b and c indicate differences in shape between Neanderthals and modern humans as shown in Fig. 4.

age); (2) centroid size (S , the extent of the landmark configuration); and (3) taxon (Neanderthal compared to modern human).

In all three analyses, only the first two statistically independent relative warps (w_1 and w_2) covary significantly with age, S and taxon, which together account for about 60% of the total shape variability (see Methods and Supplementary Information for details of relative warp analysis). There are clear patterns of covariation: w_1 describes shape variation related to size and age (Figs 1 and 2), whereas w_2 describes taxon-specific shape variation (Fig. 1). Each w_1 - w_2 coordinate in Fig. 1 corresponds to a specific morphological configuration in physical space, and each vector (arrows along statistically independent relative warps in Fig. 1) indicates a distinct spatial pattern of morphological change. The most important result of the analyses (evident from Fig. 1) is that taxon-specific differences in craniomandibular shape are present by dental stage 2 and remain subsequently unchanged during ontogeny. This indicates that the characteristic morphologies that distinguish both Neanderthals and modern humans develop before dental stage 2, during early post-natal or possibly during prenatal ontogeny. From dental stage 2 onwards, Neanderthals and modern humans follow parallel ontogenetic trajectories along the direction of w_1 (Fig. 1), demonstrating a shared spatial pattern of morphological change (Fig. 3). The different lengths of the trajectories in Fig. 1, however, indicate that there are heterochronic differences between the taxa in their post-natal ontogeny. Plots of the relationship between dental age, craniomandibular shape w_1 and size S (Fig. 2) show that although

the two taxa follow similar ontogenetic allometries (Fig. 2a), Neanderthals compared to modern humans show rate hypermorphosis (faster rates of growth and development leading to greater adult values of size and shape) during ontogeny (Fig. 2b, c). In all analyses, the fossil modern human subsample falls within the range of ontogenetic variability displayed by the extant modern human sample. In addition, the Neanderthals analysed here, which sample a long period of time¹, exhibit similar within-taxon ontogenetic variability, suggesting long-term stability of the modern human and Neanderthal patterns of ontogeny.

To further explore taxon-specific regional differences in cranial and mandibular growth, Figs 3 and 4 express statistically independent factors w_1 and w_2 as modifications of the Neanderthal/modern human consensus morphology (computed from a 5-year-old modern human specimen). Components of shape change were computed perpendicular and parallel to the craniomandibular surface and were visualized, respectively, using colours and vector fields (see Methods). Figure 3 shows the shared Neanderthal/modern human pattern of craniomandibular shape change from dental stage 2 to adulthood. The stability of this pattern over an extended phase of ontogeny (linear trajectories in Fig. 1a) indicates that the spatial distribution of areas of bone deposition/resorption and the relative rates of growth in these areas were constant in both taxa¹⁵. Patterns of Neanderthal and modern human ontogeny from dental stage 2 onwards therefore apparently derive from generally similar growth processes.

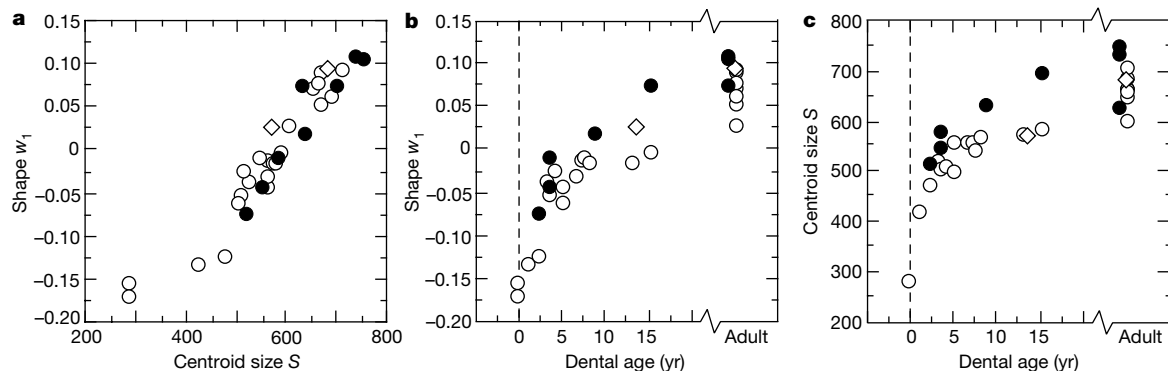


Figure 2 Correlations between shape (w_1), centroid size S and dental age (in postnatal years) for craniomandibular morphologies of Neanderthals (filled circles) and extant/fossil modern humans (open circles/diamonds, respectively). **a**, Ontogenetic allometry (w_1

versus S). **b**, Development (w_1 versus dental age). **c**, Growth (S versus dental age). Compared to modern humans, Neanderthals show rate hypermorphosis during development and growth.

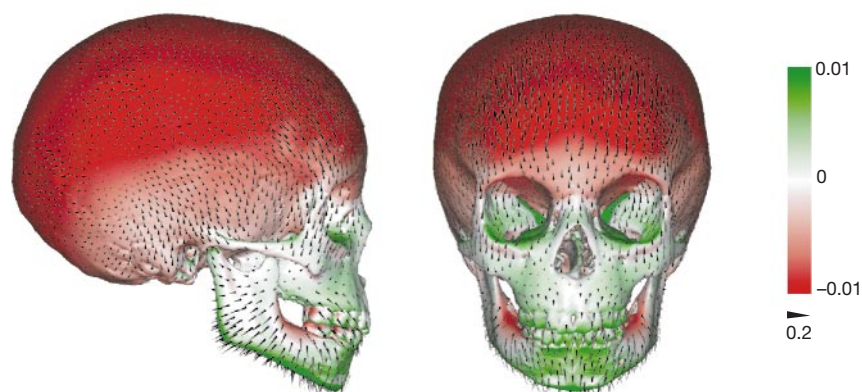


Figure 3 Patterns of shape change during postnatal development in Neanderthals and modern humans. A Neanderthal/modern human consensus skull is shown. Changes in morphology (corresponding to transformation along the arrow in Fig. 1a) are visualized with colours and vector fields (see Methods). Red and green indicates direction (inward and outward, respectively) and magnitude of shape change perpendicular to the surface

of the skull. Arrows indicate direction and magnitude of shape change tangential to the surface (all scales in units of centroid size, S). Note the projection and downward elongation of the face and mandible, as opposed to the relative 'contraction' of the cranial vault (size S is held constant).

Figure 4 shows quantitative differences between Neanderthal and modern human morphologies that are already established at dental stage 2 and that are consistent with previous characterizations of the differences between the taxa^{1,8,16–19}. Relative to modern humans, Neanderthals have a low cranial vault that is expanded posterolaterally. The boundary between superior and inferior regions of the vault (as visualized by colour in Fig. 4a–c) probably corresponds to the circumcranial reversal line that separates depository from resorptive growth fields on the internal surface of the braincase²⁰. Increased drift and displacement in the inferior region of the Neanderthal cranial vault may therefore account for many of the Neanderthal apomorphies (derived characters) found, such as a broadened temporal region, rounded lateral cranial walls¹, a more caudal position of the middle cranial fossa, an elongated foramen magnum⁸, and a large occipital squama²¹ (Fig. 4a–c). The differences in growth fields that may account for these features probably reflect taxon-specific differences in interactions between cranial base shape and brain size/shape during early development²¹.

Differential activity of growth fields may also explain many unique aspects of Neanderthal maxillofacial and mandibular morphology (Fig. 4d–f). In the fetal/early postnatal mandible, the lingual surfaces of the corpus and the external and anterior surfaces of the rami are resorptive²⁰, such that differential growth patterns affect the inclination of the rami relative to the cranial midplane and their anteroposterior position relative to the dentition. Visualization of relative warp analyses suggests that, compared to modern humans, the Neanderthal mandible was characterized by relatively less mediolateral growth and more anteroposterior growth: the rami are relatively inclined (having short condylar processes that are positioned laterally), and the corpus and rami are at a posterior position relative to the dentition (suggesting an early rather than late¹⁹ ontogenetic origin of the retromolar space^{16,17,22}). In humans, depository growth fields in the most anterior regions of the maxilla and mandible turn into resorptive fields shortly after birth, limiting facial projection during postnatal ontogeny^{20,23}. According to our data, Neanderthals follow a similar postnatal pattern. Thus,

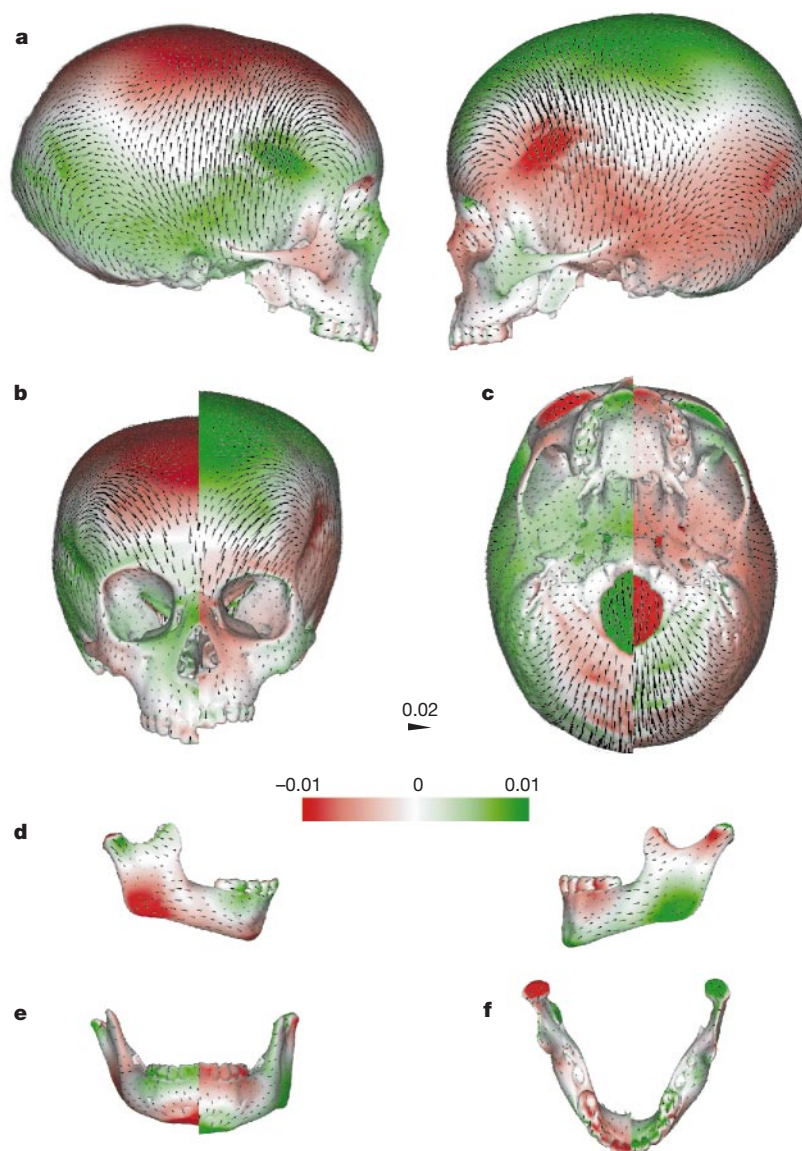


Figure 4 Shape differences between Neanderthal and modern human skulls (a–c) and mandibles (d–f). Left panels show Neanderthal morphologies; right panels show modern human morphologies (corresponding to opposite tips of the arrows in Fig. 1b, c). The patterns of colour indicate the direction and magnitude of shape difference of each taxon relative to the other and can tentatively be interpreted as representing contrasts between

modes of prenatal (and early postnatal) development. Red and green indicates shape differences perpendicular (inward and outward, respectively) to the cranial and mandibular surfaces. Arrows indicate differences tangential to surfaces (all scales in units of centroid size, *S*). An animated version (Neanderthal–modern human transition) is available as Supplementary Information.

increased rates of fetal growth and/or later reversal of growth fields may be responsible for many characteristic aspects of Neanderthal midfacial prognathism, such as expanded maxillae, a spacious nasal cavity, a receding cheek region (Fig. 4a–c), and a longer persistence of the premaxillary suture²⁴. Similar differences may generate the receding symphyseal region of the Neanderthal mandible.

The above analyses therefore indicate that most of the principal differences between Neanderthal and modern human skulls result from ontogenetically early differences in the relative timing and rates of activity at specific growth fields. These data thus support recent studies suggesting that early modifications of growth processes, notably in basicranial morphogenesis, have a principal role in generating evolutionary novelty in the hominid cranium^{21,23,25,26}. Pronounced basicranial flexion at the spheno-occipital synchondrosis before 2 yr accounts for the unique position of the modern human face beneath the anterior cranial fossa^{23,27}. It is therefore probable that the midfacial projection that is characteristic of Neanderthals^{16–18} (Fig. 4a, b) results, in part, from less basicranial flexion during early postnatal or possibly prenatal ontogeny²³. The early appearance of taxon-specific features between Neanderthals and modern humans, the morphological distinctiveness of these taxa throughout later postnatal ontogeny, and the evidence for evolutionary stasis of taxon-specific patterns of ontogeny, all support the theory^{8,12,18,25} that *Homo neanderthalensis* and *Homo sapiens* represent morphologically discrete, separate species, which belonged to distinct evolutionary lineages²⁸. It remains to be tested whether the observed pattern of evolutionary developmental diversification—prenatal divergence in conjunction with postnatal parallelism of developmental patterns—is characteristic for hominid evolution in general. □

Methods

Sample structure

The Neanderthal sample comprises 11 immature and 5 adult specimens, respectively: Amud 7 (A7; 0.5 yr); Pech de l'Azé (P; 2.2 yr); Barakai (B; 3 yr); Subalyuk 2 (S; 3.2 yr); Roc de Marsal (R; 3.5 yr); Devil's Tower (Gibraltar 2) (D; 3.5 yr); Engis 2 (E; 5.5 yr); La Quina 18 (Q; 6.5 yr); Teshik Tash (TT; 8.5 yr); La Naulette (N; 14 yr); Le Moustier 1 (M; 15 yr), all of which are immature; and Amud 1 (A1); La Ferrassie 1 (F1); Forbes' Quarry (Gibraltar 1) (G); Kebara 2 (K2); and Tabun 1 (T1), which are adult specimens. The fossil modern human sample comprises 3 specimens from the Near East: Skhul 1 (sk1; 3.5 yr); Qafzeh 11 (q11; 13.5 yr); and Qafzeh 9 (q9; adult), all of which are dated to about 100,000 yr ago¹. The extant modern humans are represented by a pooled-sex sample of 22 specimens from Europe, Africa, Asia, America and Australia. This sample includes hyper-robust specimens from high north/south latitudes (for details see Supplementary Information Table 1).

Ageing

We estimated ages at death of both Neanderthal and modern human specimens with modern standard scores of dental eruption²⁹; we used ages based on perikymata where available². Direct age estimates based on perikymata indicate that application of modern human dental scores may bias results towards higher ages in Neanderthals². As Neanderthal cranial ontogeny tends to be slightly fast relative to modern humans^{2,9,10}, dental scoring yields conservative estimates for individual ages. To distinguish early versus late phases of postnatal ontogeny, we discern between dental stage 2 (after eruption of dc (lower deciduous canine) and before full crown development of I1) and dental stage 3 (crown of I1 fully developed and before eruption of M1).

Data acquisition and fossil reconstruction

After acquisition of volume data with computer tomography, we generated three-dimensional, graphical object representations of all specimens. We reconstructed the fossil specimens by using special-purpose virtual reality tools³⁰. Briefly, filling material was removed from earlier reconstructions, the isolated fossil fragments were re-composed on the computer screen, missing parts were completed with mirror-imaged counterparts, and taphonomic deformation was tentatively corrected, using the methods and anatomical criteria described elsewhere^{4,12,14}.

Landmark data registration

Three-dimensional coordinates of the anatomical landmarks were determined on the graphical representations of the specimens. Virtual registration allows landmark data acquisition on external and internal cranial surfaces⁴. The data set used in GMM analysis comprises 51 cranial (15 midsagittal and 18 bilateral pairs) and 22 mandibular (4 midsagittal and 9 bilateral pairs) landmarks (for details see Supplementary Information Table 2). All landmarks were chosen to represent locations of between-specimen homology.

Analysis of shape variability

Relative warp analysis provides an effective means to detect and analyse patterns of correlated shape change in multi-landmark configurations. We performed relative warp analysis according to methods in ref. 6. To constrain modes of shape variation to bilateral symmetry, each specimen's landmark configuration was made symmetrical by generalized least-squares (GLS) superimposition with its mirror-imaged counterpart, and subsequent averaging. The form of each specimen is defined as the set of its three-dimensional landmark coordinates. As a linear measure of size, centroid size S is used, calculated as the square root of the sum of squared distances of all landmarks from the centre of gravity (the centroid) of the landmark configuration². GLS fitting of the size-normalized ($S = 1$) landmark configurations of all specimens yields an average configuration (the consensus) and defines a linearized Procrustes space in which the shape of each specimen is given by its linear deviation (landmark by landmark) from the consensus. Using the thin plate spline (TPS) eigenfunctions of the consensus, each specimen is then expressed as a smooth deformation of the consensus, given by its partial warp scores². An array of orthogonal basis vectors, the relative warps, were calculated to reduce the dimensionality of the resulting partial warp space. Relative warps are statistically independent factors of shape variation that account for the largest, second largest and successively smaller proportions of the total sample variance in shape (for details see Supplementary Information Table 3).

Visualization of shape change

As relative warps are TPS functions, each point in the space of relative warps (see Fig. 1) corresponds to one specific physical landmark configuration resulting from a deformation of the consensus. Similarly, the transformation between any two points in relative warp space (arrows in Fig. 1) can be expressed as a TPS function. This TPS function defines a displacement vector at every point of the cranial surface. The resulting spatial pattern of shape change is visualized as follows: for each point on the cranial surface, the displacement vector is decomposed into its normal and tangential components, relative to the local orientation of the surface. The normal component is colour coded—red and green indicates its orientation (inward and outward, respectively); colour intensity its magnitude. The tangential component is visualized as a vector field, indicating its direction and magnitude (see Figs 3 and 4).

Received 29 December 2000; accepted 26 June 2001.

1. Stringer, C. B. & Gamble, C. *In Search of the Neanderthals: Solving the Puzzle of Human Origins* (Thames and Hudson, London, 1993).
2. Krings, M. *et al.* A view of Neanderthal genetic diversity. *Nature Genet.* **26**, 144–146 (2000).
3. Ovchinnikov, I. V. *et al.* Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* **404**, 490–493 (2000).
4. Zollikofer, C. P. E., Ponce de León, M. S. & Martin, R. D. Computer-assisted paleoanthropology. *Evol. Anthropol.* **6**, 41–54 (1998).
5. Bookstein, F. L. *Morphometric Tools for Landmark Data* (Cambridge Univ. Press, Cambridge, 1991).
6. Rohlf, F. J. In *Contributions to Morphometrics* (eds Marcus, L., Bello, E. & García-Valdecasas, A.) 131–159 (Consejo superior de investigaciones científicas, Madrid, 1993).
7. Tillier, A.-M. In *The Human Revolution. Behavioural and Biological Perspectives on the Origins of Modern Humans* (eds Mellars, P. & Stringer, C. B.) 286–297 (Edinburgh Univ. Press, Edinburgh, 1989).
8. Rak, Y., Kimbel, W. H. & Hovers, E. A Neanderthal infant from Amud Cave, Israel. *J. Hum. Evol.* **26**, 313–324 (1994).
9. Dean, M. C., Stringer, C. B. & Bromage, T. G. Age at death of the Neanderthal child from Devil's Tower, Gibraltar and the implications for studies of general growth and development in Neanderthals. *Am. J. Phys. Anthropol.* **70**, 301–309 (1986).
10. Stringer, C. B., Dean, M. C. & Martin, R. D. In *Primate Life History and Evolution* (ed. DeRousseau, C. J.) 115–152 (Wiley-Liss, New York, 1990).
11. Williams, F. I. E. In *Neanderthals on the Edge* (eds Stringer, C. B., Barton, R. N. E. & Finlayson, J. C.) 257–267 (Oxbow Books, Oxford, 2000).
12. Zollikofer, C. P. E., Ponce de León, M. S., Martin, R. D. & Stucki, P. Neanderthal computer skulls. *Nature* **375**, 283–285 (1995).
13. Ponce de León, M. S. & Zollikofer, C. P. E. New evidence from Le Moustier 1: computer-assisted reconstruction and morphometry of the skull. *Anat. Rec.* **254**, 474–489 (1999).
14. Ponce de León, M. S. *Neanderthal Ontogeny: a Geometric Morphometric Analysis of Cranial Growth*. PhD thesis, Univ. Zürich (2000).
15. O'Higgins, P. & Jones, N. Facial growth in *Cercopithecus torquatus*: an application of three-dimensional geometric morphometric techniques to the study of morphological variation. *J. Anat.* **193**, 251–272 (1998).
16. Rak, Y. The Neanderthal: a new look at an old face. *J. Hum. Evol.* **15**, 151–164 (1986).
17. Trinkaus, E. The Neanderthal face: evolutionary and functional perspectives on a recent hominid face. *J. Hum. Evol.* **16**, 429–443 (1987).
18. Schwartz, J. H. & Tattersall, I. Significance of some previously unrecognized apomorphies in the nasal region of *Homo neanderthalensis*. *Proc. Natl Acad. Sci. USA* **93**, 10852–10856 (1996).
19. Rosas, A. Occurrence of Neanderthal features in mandibles from the Atapuerca-SH site. *Am. J. Phys. Anthropol.* **114**, 74–91 (2001).
20. Enlow, D. H. *Facial Growth* 3rd edn (Saunders, Philadelphia, 1990).
21. Lieberman, D. E., Pearson, O. M. & Mowbray, K. M. Basicranial influence on overall cranial shape. *J. Hum. Evol.* **38**, 291–315 (2000).
22. Franciscus, R. G. & Trinkaus, E. Determinants of retromolar space presence in Pleistocene *Homo* mandibles. *J. Hum. Evol.* **28**, 577–595 (1995).
23. Lieberman, D. E. In *Development, Growth and Evolution* (eds O'Higgins, P. & Cohn, M. J.) 85–122 (Linn. Soc. Lond., London, 2000).
24. Maureille, B. & Bar, D. The premaxilla in Neanderthal and early modern children: ontogeny and morphology. *J. Hum. Evol.* **37**, 137–152 (1999).

25. Lieberman, D. E. Sphenoid shortening and the evolution of modern human cranial shape. *Nature* **393**, 158–162 (1998).
26. Lieberman, D. E., Ross, C. F. & Ravosa, M. J. The primate cranial base: ontogeny, function, and integration. *Yb. Phys. Anthropol.* **43**, 117–169 (2000).
27. Lieberman, D. E. & McCarthy, R. C. The ontogeny of cranial base angulation in humans and chimpanzees and its implications for reconstructing pharyngeal dimensions. *J. Hum. Evol.* **36**, 487–517 (1999).
28. Tattersall, I. & Schwartz, J. H. Morphology, paleoanthropology and Neanderthals. *New Anat.* **253**, 113–117 (1998).
29. Ubelaker, D. H. *Human Skeletal Remains. Excavation, Analysis, Interpretation* (Chicago Univ. Press, Chicago, 1978).
30. Zollikofer, C. P. E. & Ponce de León, M. S. Tools for rapid prototyping in the biosciences. *IEEE Comp. Graph. Appl.* **15**, 48–55 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We are grateful to R. D. Martin and P. Stucki for support of our research. C. B. Stringer's continuous help and advice is gratefully acknowledged. We thank C. Howell, J.-J. Jaeger, D. Lieberman, J. Schwartz and C. Pryce for valuable comments on an earlier version of this paper. We appreciate the help of the following curators in providing access to fossil specimens: J.-J. Cleyet-Merle, J.-M. Cordy, V. Kharitonov, M. Marinot, F. Menghin, R. Orban, I. Pap, Y. Rak, C. B. Stringer, B. Vandermeersch. We also thank the radiologists, physicists and technicians engaged in fossil computer tomography scanning: E. Berenyi, G. Bijl, P. Dondelinger, V. Dousset, W. Fuchs, S. Louryan, V. Makarenko, U. Shreter, N. Strickland and C. L. Zollikofer. This work was supported by the Swiss National Science Foundation and a habilitation grant of the Canton of Zurich.

Correspondence and requests for materials should be addressed to C.Z. (e-mail: zolli@ifi.unizh.ch).

Habitat structure and population persistence in an experimental community

Stephen P. Ellner^{*}, Edward McCauley[†], Bruce E. Kendall[‡], Cheryl J. Briggs[§], Parveiz R. Hosseini^{||}, Simon N. Wood[¶], Arne Janssen[#], Maurice W. Sabelis^{*}, Peter Turchin[☆], Roger M. Nisbet^{||} & William W. Murdoch^{||}

^{*} Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853-2701, USA

[†] Ecology Division, Department of Biological Sciences, University of Calgary, Calgary T2N 1N4, Canada

[‡] Donald Bren School of Environmental Science and Management; and

^{||} Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, California 93106, USA

[§] Department of Integrative Biology, University of California, Berkeley, California 94720, USA

[¶] School of Mathematical and Computation Sciences, University of St Andrews, Fife, Scotland KY16 9SS, UK

[#] Institute for Biodiversity and Ecosystem Dynamics, PO Box 94084, 1090 GB Amsterdam, The Netherlands

[☆] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269, USA

Understanding spatial population dynamics is fundamental for many questions in ecology and conservation^{1–4}. Many theoretical mechanisms have been proposed whereby spatial structure can promote population persistence, in particular for exploiter–victim systems (host–parasite/pathogen, predator–prey) whose interactions are inherently oscillatory and therefore prone to extinction of local populations^{5–11}. Experiments have confirmed that spatial structure can extend persistence^{11–16}, but it has rarely been possible to identify the specific mechanisms involved. Here we use a model-based approach to identify the effects of spatial population processes in experimental systems of bean plants

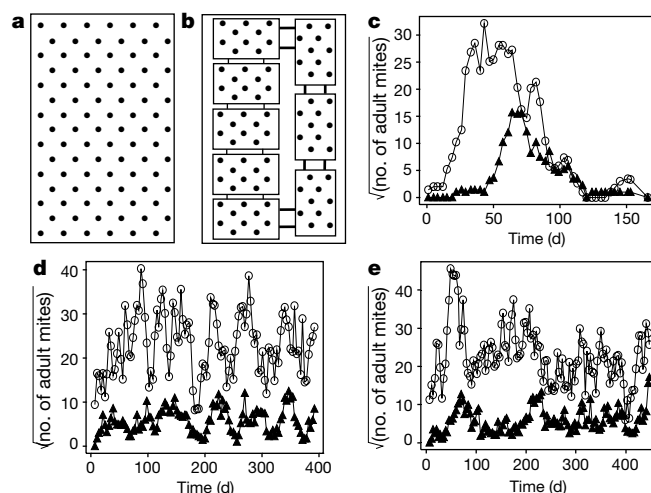


Figure 1 Experimental layouts and results. **a**, The single-island system consisted of a Styrofoam sheet with 90 embedded plants (filled circles) floating in a shallow tray of water. **b**, The metapopulation subdivided the sheet into 8 islands (10 plants per island) connected by cork bridges, with the space for 10 plants being lost. Replicate systems were housed simultaneously in the same environmental chamber, and given identical initial inoculations of mites. **c**, Fluctuations in total density of prey (open circles) and predatory (filled triangles) mites in the single-island experiment. **d, e**, Fluctuations in total density of prey (open circles) and predatory (filled triangles) mites in the two replicates of the metapopulation experiment.

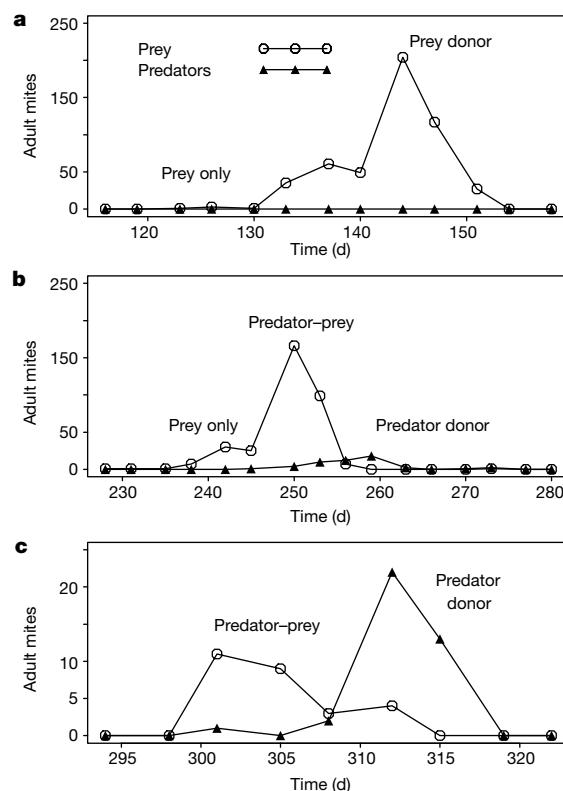


Figure 2 Examples of mite population dynamics on a single plant, from run B of the metapopulation experiment. **a**, A prey outbreak (plant 1, island 5) that was not discovered by predators. The three successively larger peaks in the prey density (days 126, 137 and 144) are the original colonizers, their offspring (counted when they become adults at age ~10 d), and offspring of the offspring. The rapid collapse of the outbreak is primarily due to emigration after exhaustion of the resource. **b**, A prey outbreak (plant 1, island 2) colonized by predators after several on-plant prey generations. Predators arrived too late to prevent growth of the prey population, and the outbreak terminated through exhaustion of the resource and prey emigration. **c**, A prey outbreak (plant 8, island 2) colonized by predators when prey densities were still low. The outbreak terminated through predators consuming all prey and then emigrating.

(*Phaseolus lunatus*), herbivorous mites (*Tetranychus urticae*) and predatory mites (*Phytoseiulus persimilis*). On isolated plants, and in a spatially undivided experimental system of 90 plants, prey and predator populations collapsed; however, introducing habitat structure allowed long-term persistence. Using mechanistic models, we determine that spatial population structure did not contribute to persistence, and spatially explicit models are not needed. Rather, habitat structure reduced the success of predators at locating prey outbreaks, allowing between-plant asynchrony of local population cycles due to random colonization events.

Plants (*P. lunatus*) were grown in individual pots in arrays that were initially free of mites¹⁵ (Fig. 1). Experiments began by placing individual prey mites (*T. urticae*) onto several plants, followed by predatory mites (*P. persimilis*) on the same plants. Mite populations on an individual plant collapse either through the prey exhausting the resource or by predators colonizing the plant, consuming all prey and then starving or emigrating¹⁷ (Fig. 2). The individual plant dynamics did not differ between the single-island (unstructured habitat) and metapopulation (subdivided habitat) systems (Fig. 1a, b), therefore the differing system dynamics must result from effects of habitat structure. The factors are present in the metapopulation system for two distinct but non-exclusive stabilizing mechanisms.

The first mechanism is that spatial subdivision produces differing dispersal scales in predators and prey. In the single-island system, both species dispersed throughout the system, but in the metapopulation system only predators dispersed throughout the system while prey mainly colonized plants on the same or nearby

islands¹⁸. Such differences in dispersal can create spatial patterns where different locations act as source habitat for prey and predators, and total densities are stabilized^{19,20}. We tested for such clumped distributions using Moran's *I*, and found significant ($P < 0.05$) spatial correlation ($I > 0$) among plants on the same island for both mite species on most census dates (85 and 81% for prey; 93 and 92% for predators, in the two metapopulation experiments). The second mechanism is that spatial subdivision reduces the discovery rate of prey outbreaks by predators. Nearly 90% of prey outbreaks were attacked by predators in the single-island experiments, but only 68 and 78% in the two metapopulation experiments¹⁸. This mechanism has two components, neither involving spatial pattern: reduced average predator success, and the resulting between-plant asynchrony, meaning that not all plants are simultaneously exploited by prey and not all prey outbreaks are simultaneously discovered and extinguished. Hence a non-spatial model for total numbers of plants occupied by each species can describe the system²¹.

To determine which mechanism accounts for the observed dynamics, we developed and tested models that allowed us to simulate experiments that would be infeasible in the actual systems. The models consist of colonization probabilities, and nonlinear stage-structure models for the predator-prey-plant interactions occurring on the plant. We assigned parameters to the on-plant interaction models (Box 1) using independent data^{22,23}; these were identical for the metapopulation and single-island systems. The equations for colonization probabilities (Box 2) give each plant's daily chance of colonization as a function of the abundance and

Box 1

Structured population models for mite dynamics on a plant

We combined a stochastic description of plant colonizations (Box 2) with a deterministic growth model for mite populations on occupied plants. The stochasticity of individual mites colonizing suitable plants is essential for generating between-plant asynchrony in the model, but the on-plant dynamics involve many individuals and can be modelled deterministically^{17,23}.

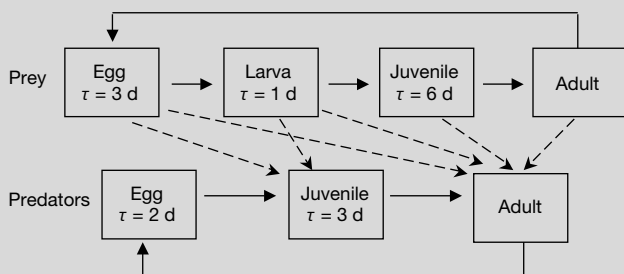
We used nonlinear, stage-structured²⁶ matrix models²⁷ with a 1-d time step for the mite populations on each occupied plant. We classified mites by life stage and assigned stage-specific vital rates (described below) derived from empirical age- and sex-specific vital rates^{22,23}. Male adults are much smaller than female adults in both species; we gave male adults the same feeding and survival rates as juvenile females (who are similar in size), and assumed a fixed male:female ratio in all stages (1:4 in predators; 1:3 in prey) rather than explicitly tracking males.

Because pre-adult prey require much less food and rarely experience food shortages, we assumed predation to be the only cause of pre-adult

prey mortality. It is a reasonable approximation for prey that daily fecundity is 9 d^{-1} at maturation, and egg production (survival to age $\times$ fecundity at age a) subsequently declines by $10\% \text{ d}^{-1}$. We therefore modelled adult prey as having a constant fecundity at 9 d^{-1} , and $10\% \text{ d}^{-1}$ mortality in the absence of predators. Similarly, adult predators were assumed to have constant fecundity (3.2 d^{-1}) and mortality rates (0.04 d^{-1}) when prey are not limiting.

Predator development from egg to adult requires the consumption of 15 prey eggs, which we modelled as a requirement of 5 eggs d^{-1} ; adult predators require 12 eggs d^{-1} . On the basis of body mass we modelled prey larvae as equivalent to 1 egg, juveniles and adult males as 12/7 eggs, and adult females as 6 eggs²². Predators had a strict feeding preference for prey at younger life stages^{23,28}. We assumed predator food requirements and preferences were satisfied unless limited by prey abundance on the plant. Predator fecundity scaled down linearly from its maximum value when food requirements are met, to zero when no feeding occurs. This is a good description for the number of female eggs laid (the fraction of male eggs is higher when prey are scarce²⁹). Similarly, we assumed predator survival decreased linearly in proportion to unmet food requirements, to a minimum of 0.2 d^{-1} when no feeding occurs; most mortality occurs when food-deprived predators emigrate and search for another plant with prey.

We assumed that adult, female prey mites damage plant leaf area at a constant daily rate per individual. When the area damaged exceeds a threshold, egg laying ceases, adults begin to emigrate at an increasing rate, and juveniles experience increasing mortality and emigrate when they mature. The emigration model was assigned parameters from observations of isolated plants with only prey mites present: twice-weekly counts of adults on the plant, adults recaptured near the plant, and the percentage of damaged leaf area (A.J. and E. van Gool, unpublished data). The start of emigration coincided with the peak in prey density. We therefore set the damage threshold at the estimated average number of mite days between initial colonization and peak density on plants where prey outbreaks went undetected from the single-island and metapopulation experiments.



Box 1 Figure

Life stages of prey and predatory mites in the model. Solid arrows indicate survival and fecundity, dashed arrows indicate predation (other mortality losses are omitted for clarity). We assumed constant durations τ for pre-adult stages²². Adult stages have indefinite duration, with a constant background mortality rate (0.1 d^{-1} for prey; 0.04 d^{-1} for predators). The resulting models are coupled, nonlinear Leslie matrices of sizes 11×11 and 6×6 for prey and predators, respectively.

Box 2

Colonization probability models

The colonization models are regression models for the conditional probability of an available plant being colonized, given the current and recent mite abundances on each plant, and the state of each plant. Plant states are as indicated in Fig. 2: empty; prey only; prey donor (plant is exhausted and prey are emigrating); predator–prey; and predator donor (all prey consumed and predators emigrating). Either of the donor states is followed by the refractory state, ending with replacement by a fresh plant (following the experimental protocol. Small prey outbreaks (number of adults is always ≤ 10) were not scored as colonization events or prey-only state; we interpret these as mites exploring a plant without settling. However one adult female of each species was sufficient for the predator–prey state, because a few predators arriving early in an outbreak can consume prey before they mature.

We used logistic regression with stepwise and backward variable selection (as in ref. 18) to identify significant 'risk factors' affecting colonization. Plants were classified by distance from the focal plant (single-island: nearest neighbour, the next nearest neighbour, the next, next-nearest neighbour, and non-neighbour; metapopulation: same island, adjacent island, or non-adjacent island; adding within-island distance classes did not improve the final model). The potential risk

factors were the mean conspecific densities within each distance class, the decrease in density between current and previous censuses on donor state plants for the species (an indicator of emigration rate), and prey density on the focal plant for predator colonization. We pooled the two metapopulation experiments and did not consider time-lagged densities because mites tend to find new plants quickly or not at all. As an example, the fitted predator colonization probability (between now and the next census) for a prey-only plant in the metapopulation is:

$$\begin{aligned} \log[P/(1-P)] = & -2.7 + 0.074 \text{ (prey density on plant)}^{0.5} \\ & + 0.024 \text{ (total predator density on same island)} \\ & + 0.039 \text{ (total predator decrease on same-island} \\ & \text{predator donor plants)} \\ & + 0.021 \text{ (total predator decrease on adjacent-island} \\ & \text{predator donor plants)} \\ & + 0.011 \text{ (total predator decrease on} \\ & \text{non-adjacent-island predator donor plants)} \end{aligned}$$

spatial distribution of predator and prey mites; these were estimated from the modelled experiments and differed between the metapopulation and single-island systems. However, neither mean colonization rates (number of plants colonized each day) nor their patterns of temporal variation were fitted to the data. Rather, they are predicted by the model from the interaction between on-plant dynamics and the equations for colonization probabilities.

The model dynamics match the experimental results. Nearly all model runs of the single-island system exhibit rapid increase of prey and then predators, followed by extinction (Fig. 3a, b). Extinction of both species occurred within 1 yr in over 99% of 1,000 simulation runs. Replicate runs are highly variable because small chance differences in the success of the predators at detecting the initial prey outbreaks have a large effect on the peak densities reached by prey and then predators. Similarly, the single-island experiment (Fig. 1c) persisted longer than typical model runs (Fig. 3a, b) owing

to two late prey outbreaks that started at about day 50, but escaped predation until days 65 and 78.

The metapopulation model, although differing only in its equations of colonization probability, showed long-term persistence much more frequently (Table 1 and Fig. 3c, d), corresponding to the experiments (Fig. 1d, e). Measures of temporal averages (Table 1) and patterns of temporal variation (Fig. 4) were close to those for the experiments. The model also produced spatial correlations similar to those in the experiments (within-island I was significantly positive on 63–90% of sample dates (mean = 74%) for prey; 83–97% of sample dates (mean = 89%) for predators, in 25 model runs with both species persisting). These are strong tests because we did not use any of these summary statistics in assigning parameters to the models. However, the model over-predicts slightly the number of plants occupied by prey while under-predicting the prey colonization rate; the coefficients of variation of population fluctuations were low (possibly because the on-plant population

Table 1 Summary statistics for metapopulation experiments and models with and without spatial pattern at different scales

	Experiments (run1, run2)	Metapopulation (spatial model)	Island shuffle	Plant scramble	Global dispersal	Blur plants on island
Prey						
Mean no.	610, 587	696 (613, 782)	695 (612, 786)	688 (601, 778)	650 (562, 740)	260 (112, 420)
CV (%)	55, 68	48 (41, 57)	48 (41, 56)	48 (41, 57)	51 (43, 59)	205 (172, 250)
Plants occupied	7.8, 8.1	8.7 (7.7, 9.6)	8.6 (7.7, 9.6)	8.6 (7.6, 9.6)	8.2 (7.2, 9.2)	4.1 (1.9, 6.4)
Colonization rate (plants d ⁻¹)	0.83, 0.84	0.61 (0.56, 0.67)	0.61 (0.56, 0.67)	0.61 (0.56, 0.67)	0.60 (0.55, 0.66)	0.43 (0.36, 0.48)
Predators						
Mean no.	42, 47	44 (34, 54)	44 (34, 54)	45 (36, 56)	47 (36, 58)	61 (37, 83)
CV (%)	87, 104	74 (60, 89)	74 (61, 88)	73 (60, 88)	74 (60, 88)	223 (184, 275)
Plants occupied	7.0, 7.8	6.6 (5.7, 7.7)	6.6 (5.8, 7.6)	6.8 (5.8, 7.8)	6.9 (6.0, 7.9)	9.1 (6.8, 10.8)
Colonization rate (plants d ⁻¹)	0.52, 0.59	0.48 (0.42, 0.54)	0.48 (0.42, 0.54)	0.49 (0.43, 0.55)	0.50 (0.44, 0.56)	0.83 (0.63, 1.09)
Persistence						
Prey (% of runs)	yes, yes	100	100	100	100	1.8
Both species (% of runs)	yes, yes	81	79	81	80	0.08

Values for the models are mean values (5th and 95th percentiles in parentheses) from 1,000 runs (50,000 for plant blur), averaging over runs in which both species persisted for 400 d. Model runs in which one or both species go extinct are not comparable to the experiments, in which both species persisted. Data from the first 30 d of the experiments and simulations were omitted to eliminate transients. Plants were scored as occupied by prey (for model output and experimental data) only when ten or more adults were present, to eliminate the frequent counts of a few stray prey in the experimental systems not associated with a prey outbreak on a plant. CV, coefficient of variation.

models are deterministic); and the time lags between changes in prey abundance and the resulting changes in predator abundance and colonization rate were under-predicted by roughly 3 d.

The differing temporal autocorrelations between the two metapopulation experiments¹⁵ (Fig. 4c–f) are also consistent with the model. The first wave of predator colonizations was much smaller in the second experiment (Fig. 1e) than in the first (Fig. 1d), resulting in higher prey densities and a deeper subsequent crash by both species. This pattern also occurred in about 15% of model runs where initial predator growth was restricted.

We determined which aspects of spatial dynamics were responsible for persistence in the metapopulation by systematically eliminating different aspects of spatial pattern in simulations. Under the first mechanism, eliminating spatial pattern at the island and metapopulation levels should produce dynamics similar to those observed in the single-island system. Under the second mechanism, reducing local between-plant asynchrony should have this effect. In 'island shuffling' simulations, the sequence of islands in the ring (Fig. 1b) was randomly shuffled each day, but plants remained on their home island. This preserves within-island correlation but eliminates pattern at the metapopulation level. In 'plant scrambling'

simulations the location of all plants was randomized each day, and in 'global dispersal' simulations, all colonization probabilities were calculated from system-wide average densities rather than local densities. These simulations eliminate spatial pattern at the island and metapopulation level, but retain between-plant asynchrony. In 'plant blurring' simulations, the age-specific mite densities on each plant were replaced daily by the average density over all plants on the same island. This eliminates within-island asynchrony but preserves pattern at the metapopulation level.

Island shuffling, plant scrambling and global dispersal had only small effects on the dynamics (Table 1, Figs 3 and 4). Plant scrambling eliminated within-island autocorrelations because each island is reassembled daily at random, but autocorrelations of total densities were unaffected (Fig. 4). In contrast, plant blurring produced extinction similar to the single-island experiment (Fig. 3i, j) with predators finding all prey outbreaks because of the rapid within-island dispersal. The reduced system size with blurring (8 islands versus 80 plants) still allows enough asynchrony for prey persistence in the absence of predators (81% of 1,000 runs).

These results indicate that the first stabilizing mechanism is not responsible for long-term persistence in the metapopulation

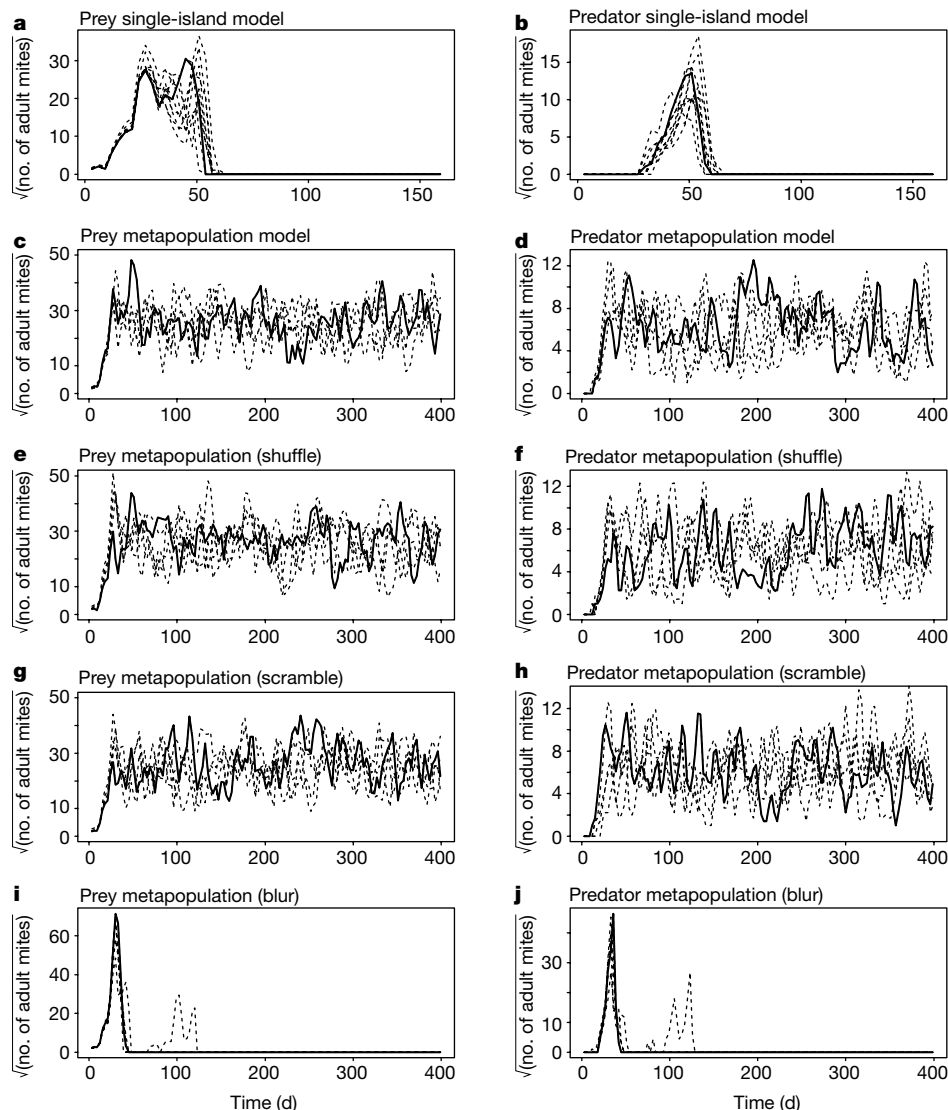


Figure 3 Total prey and predator mite densities in consecutive replicate runs of the simulation models. The first run is plotted as a solid line, the rest as dashed lines. **a, b**, Single-island model. **c, d**, Metapopulation model. **e, f**, Metapopulation model with

island shuffling. **g, h**, Metapopulation model with plant scrambling. **i, j**, Metapopulation model with plant blurring; note the change in scale. For panels **c–h** we show only runs in which both species persist to day 400, which is a typical result (Table 1).

system, and are consistent with the second mechanism. If the second mechanism caused persistence, decreasing the predators' success at detecting prey outbreaks in the single-island should result in persistence. We tested this prediction in the single-island model by reducing daily probability of predators finding each prey-only plant by a factor P ($0.05 < P < 1$). The probability of both species persisting for 360 d was highest when about 25% of prey outbreaks escape detection (88% of 1,000 runs at $P = 0.1$), which is similar to the actual values in the metapopulation experiments (22 and 32%). Conversely, the metapopulation model persisted less often when predator colonization probabilities were increased.

Thus, the dynamics of the systems results mainly from the average probability that prey mites find unoccupied plants, the average probability that predators find a prey outbreak, and the stochasticity of individual colonization events. The systems can be modelled successfully without regard to the spatial location of plants, as in classical models for metapopulation²⁴ and host–parasitoid dynamics⁵. A preliminary 'metapopulation swapping' experiment

is consistent with this conclusion (Supplementary Information).

Comparison of alternative mechanistic models is a powerful approach for testing hypotheses regarding processes responsible for patterns in ecological dynamics. By comparing models with different assumptions about the effects of habitat subdivision, we have shown that the one essential process allowing persistence of the metapopulation was one of the earliest and simplest hypotheses about spatial dynamics: isolation by distance^{9,21,24,25}. Habitat subdivision increased the effective distance between plants from the viewpoint of the predators, giving prey a moving refuge from their enemies. □

Methods

Experimental design

We counted adult female mites on each plant twice a week. After all mites had left a plant (three consecutive samples without prey or four without predators, whichever occurred last), it was replaced with a 1-week-old plant pruned to two leaves. We regularly pruned plants and replaced uncolonized plants with 1-week-old plants, to standardize conditions and eliminate direct connections among plants. Mites dispersed only by walking from one plant to another. Because the bridges between islands in the metapopulation were positioned below the island rim, mites often missed these connections, reducing inter-island movements. See ref. 15 for further experimental details. Initial prey inoculations in the models were identical to the experiments, matching the day, plant and number of individuals added. Predator inoculations in the models corresponded in the same way to successful predator inoculations in the experiments (those where the founding predator reproduced and established a local population).

Received 27 March; accepted 20 June 2001.

1. Tilman, D. & Kareiva, P. (eds) *Spatial Ecology: the Role of Space in Population Dynamics and Interspecific Interactions* (Princeton Univ. Press, Princeton, 1997).
2. Dieckmann, U., Law, R. & Metz, J. A. J. (eds) *The Geometry of Ecological Interactions: Simplifying Spatial Complexity* (Cambridge Univ. Press, Cambridge, 2000).
3. Hanski, I. A. & Gilpin, M. E. (eds) *Metapopulation Biology: Ecology, Genetics, and Evolution* (Academic, San Diego, 1997).
4. Levin, S. A., Grenfell, B., Hastings, A. & Perelson, A. S. Mathematical and computational challenges in population biology and ecosystems science. *Science* **275**, 334–343 (1997).
5. May, R. M. Host–parasitoid systems in patchy environments: a phenomenological model. *J. Anim. Ecol.* **47**, 833–843 (1978).
6. Sabelis, M. W. & Dieckmann, O. Overall population stability despite local extinction: the stabilizing effect of prey dispersal from predator-invaded patches. *Theor. Popul. Biol.* **34**, 169–176 (1988).
7. Hassell, M. P. & May, R. M. Spatial heterogeneity and the dynamics of parasitoid–host systems. *Ann. Zool. Fenn.* **25**, 55–61 (1988).
8. Hassell, M. P., Comins, H. N. & May, R. M. Spatial structure and chaos in insect population dynamics. *Nature* **353**, 255–258 (1991).
9. de Roos, A. M., McCauley, E. & Wilson, W. Mobility versus density limited predator–prey dynamics on different spatial scales. *Proc. R. Soc. Lond. B* **246**, 117–122 (1991).
10. Sabelis, M. W., Dieckmann, O. & Jansen, V. A. A. Metapopulation persistence despite local extinction—predator–prey patch models of the Lotka–Volterra type. *Biol. J. Linn. Soc.* **42**, 267–283 (1991).
11. Murdoch, W. W. Population regulation in theory and practice. *Ecology* **75**, 271–287 (1994).
12. Huffaker, C. B., Shea, K. P. & Herman, S. G. Experimental studies on predation: complex dispersion and levels of food in an acarine predator–prey interaction. *Hilgardia* **34**, 305–330 (1963).
13. van de Klashorst, G., Readshaw, G. L., Sabelis, M. W. & Lingeman, R. A demonstration of asynchronous local cycles in an acarine predator–prey system. *Exp. Appl. Acarol.* **14**, 185–199 (1992).
14. Holyoak, M. & Lawler, S. P. Persistence of an extinction prone predator–prey interaction through metapopulation dynamics. *Ecology* **77**, 1867–1879 (1996).
15. Janssen, A., van Gool, E., Lingeman, R., Jacas, J. & van de Klashorst, G. Metapopulation dynamics of a persisting predator–prey system in the laboratory: time-series analysis. *Exp. Appl. Acarol.* **21**, 415–430 (1997).
16. Holyoak, M. Habitat patch arrangement and metapopulation persistence of predators and prey. *Am. Nat.* **156**, 378–389 (2000).
17. Pels, B. & Sabelis, M. W. Local dynamics, overexploitation and predator dispersal in an acarine predator–prey system. *Oikos* **86**, 573–583 (1999).
18. McCauley, E. *et al.* Inferring colonization processes from population dynamics in spatially structured predator–prey systems. *Ecology* **81**, 3350–3361 (2000).
19. McCauley, E., de Roos, A. M. & Wilson, W. Dynamics of age- and spatially-structured predator–prey interactions: individual based models and population level formulations. *Am. Nat.* **142**, 412–442 (1993).
20. de Roos, A. M., McCauley, E. & Wilson, W. Pattern formation and the spatial scale of interactions between predators and their prey. *Theor. Popul. Biol.* **53**, 108–130 (1998).
21. Nisbet, R. M. & Gurney, W. S. C. *Modelling Fluctuating Populations* Ch. 10 (Wiley, New York, 1982).
22. Helle, W. & Sabelis, M. W. (eds) *Spider Mites, their Biology, Natural Enemies, and Control* (Elsevier, Amsterdam, 1985).
23. Sabelis, M. W. & van der Meer, J. in *Dynamics of Physiologically Structured Populations* (eds Metz, J. A. J. & Dieckmann, O.) 322–344 (Springer, New York, 1986).
24. Hastings, A. Spatial heterogeneity and stability of predator–prey systems. *Theor. Popul. Biol.* **12**, 37–48 (1977).
25. Wright, S. Isolation by distance. *Genetics* **28**, 114–138 (1943).
26. Gurney, W. S. C., Nisbet, R. M. & Lawton, J. H. The systematic formulation of tractable single-species population models incorporating age structure. *J. Anim. Ecol.* **52**, 479–495 (1983).

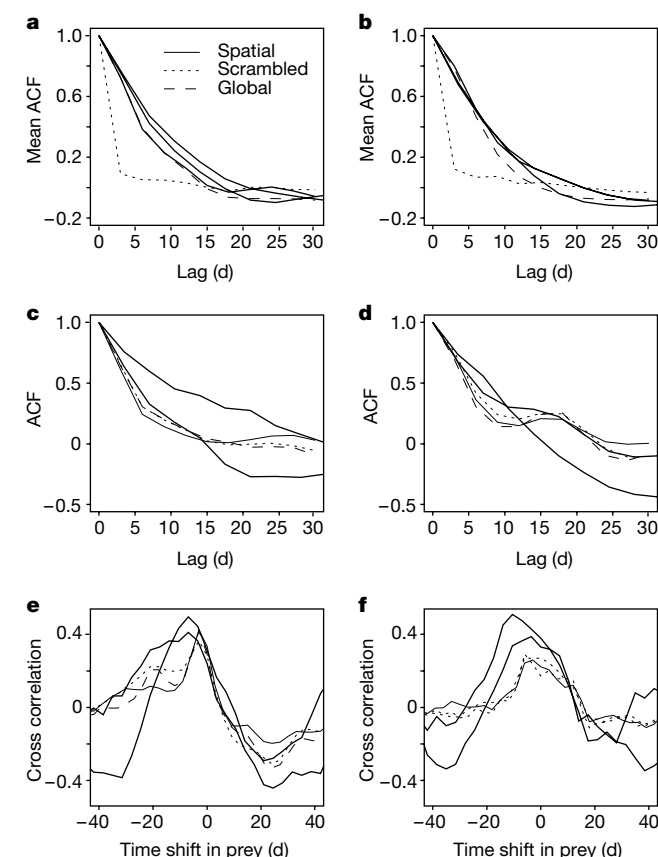


Figure 4 Summary measures of temporal patterns in model output and experimental results. In all panels the thick solid curves are the two metapopulation experiments, the thin solid curve is the metapopulation model, the dotted curve is the metapopulation model with plant scrambling (plant locations randomized each day), and the dashed curve is the model with global dispersal among plants irrespective of location. Curves for model output are the average over 10 replicate runs in which both species persist for 400 d. Correlations were calculated on square-root-transformed population counts; counts of colonization events were not transformed. **a, b**, Average across islands of the autocorrelation function (ACF) for the time series of total density on an island (8 series each for prey **a**) and predator **b**). **c, d**, Autocorrelation function of the system-wide total prey **c**) and predator **d**) densities. **e**, Cross correlation between system-wide total prey density at time $(t + L)$ and total predator density at time t , as a function of the time shift L . **f**, Cross correlation between total prey density at time $(t + L)$ and the predator colonization rate (plants d^{-1}) at time t , as a function of the time shift L . Plant scrambling has large effects at the island level (**a** and **b**) but has only small effects at the system-wide level (**c–f**).

27. Caswell, H. *Matrix Population Models: Construction, Analysis, and Interpretation* (Sinauer Associates, Sunderland, Massachusetts, 2001).
28. Sabelis, M. W. How to analyze prey preference when prey density varies? A new method to discriminate between effects of gut fullness and prey type composition. *Oecologia* **82**, 289–298 (1990).
29. Sabelis, M. W. & Nagelkerke, C. J. Sex allocation strategies of pseudoarrhenotokous phytoseiid mites. *Neth. J. Zool.* **37**, 117–136 (1987).

Supplementary information is available at Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank E. van Gool for assistance with the experiments. This work was undertaken as part of the Working Group on Complex Population Dynamics at the National Center for Ecological Analysis and Synthesis, a centre funded by the US National Science Foundation, University of California–Santa Barbara, and the State of California. A.J. and M.W.S. carried out experiments; E.M., B.E.K., C.J.B. and S.N.W. conceived and fitted colonization models; S.P.E., S.N.W., P.R.H., A.J., P.T., R.M.N. and W.W.N. conceived and fitted structured population models; S.P.E., P.R.H., S.N.W. and C.J.B. implemented the models; S.P.E., E.M. and B.E.K. carried out comparisons of models with data; and S.P.E., E.M., A.J., M.W.S. and P.R.H. prepared the original manuscript.

Correspondence and requests for materials should be addressed to S.P.E. (e-mail: spe2@cornell.edu).

The end of world population growth

Wolfgang Lutz*, Warren Sanderson*† & Sergei Scherbov‡

* International Institute for Applied Systems Analysis, Schlossplatz 1, A-2361 Laxenburg, Austria

† Departments of Economics and History, State University of New York at Stony Brook, New York 11794-4384, USA

‡ University of Groningen, PO Box 800, NL-9700 AV Groningen, The Netherlands

There has been enormous concern about the consequences of human population growth for the environment and for social and economic development. But this growth is likely to come to an end in the foreseeable future. Improving on earlier methods of probabilistic forecasting¹, here we show that there is around an 85 per cent chance that the world's population will stop growing before the end of the century. There is a 60 per cent probability that the world's population will not exceed 10 billion people before 2100, and around a 15 per cent probability that the world's population at the end of the century will be lower than it is today. For different regions, the date and size of the peak population will vary considerably.

Figure 1 shows the probability that the world population size would reach a peak at or before any given year. It indicates that there is around a 20 per cent chance that the peak population would be reached by 2050, around a 55 per cent chance that it would be reached by 2075, and around an 85 per cent chance that it would be reached by the end of the century.

There is around a 75 per cent chance that the peak population of the European portion of the former USSR has already been reached in 2000, an 88 per cent probability that it will be reached by 2025, and over a 95 per cent chance by the end of the century. For the China region, the probability of reaching a peak within the next two decades is still low owing to its relatively young age structure. By 2040 the probability becomes greater than half. In sub-Saharan Africa, despite the prevalence of HIV, there is a low probability of peaking before the middle of the century. The probability reaches 25 per cent by 2070, 50 per cent by 2085, and almost 75 per cent by 2100, owing to assumed reductions in fertility.

Figure 2 shows the distribution of simulated world population sizes over time. The median value of our projections reaches a peak around 2070 at 9.0 billion people and then slowly decreases. In 2100,

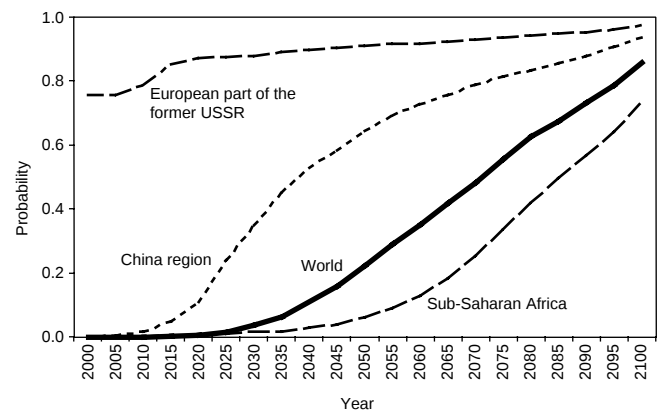


Figure 1 Forecasted probability that population will start to decline at or before the indicated date.

the median value of our projections is 8.4 billion people with the 80 per cent prediction interval bounded by 5.6 and 12.1 billion. The medium scenario of the most recent United Nations long-range projection² is inserted in Fig. 2 as a white line. It is almost identical to our median until the middle of the century, but is higher thereafter owing to the United Nations assumption of universal replacement-level fertility, that is two surviving children per woman.

Table 1 shows the median population sizes and associated 80 per cent prediction intervals for the world and its 13 regions, indicating major regional differences in the paths of population growth. While over the next two decades the medians are already declining in eastern Europe and the European portion of the former Soviet Union, the populations of north Africa and sub-Saharan Africa are likely to double, even when we take into account the uncertainty about future HIV trends.

The China region and the South Asia region, which have approximately the same population size in 2000, are likely to follow very different trends. Owing to an earlier fertility decline, the China region is likely to have around 700 million fewer people than the South Asia region by the middle of the century. This absolute difference in population size is likely to be maintained over the entire second half of the century and illustrates the strong impact of the timing of fertility decline on eventual population size³.

Our findings concerning the timing of the end of world population growth are robust to plausible changes in parameter assumptions. A detailed sensitivity analysis is provided as Supplementary Information. The forecasts of the World Bank, the US Census Bureau, and the medium variant of the United Nations^{2,4,5} are based on independent assumptions; the median trajectory of our world forecasts is almost identical to these up until 2045. Of these three forecasts, only the UN long-range projections provide scenarios of the world's population to the end of the century. If we define the end of population growth slightly less literally, and take it to correspond with annual population growth of one-tenth of one per cent or less, the United Nations medium projection also shows the end of population growth during the second half of the century. Their medium scenario predicts that world population growth will first fall below one-tenth of one per cent at around 2075.

A stabilized or shrinking population will be a much older population. At the global level the proportion above age 60 is likely to increase from its current level of 10 per cent to around 22 per cent in 2050. This is higher than it is in western Europe today. By the end of the century it will increase to around 34 per cent, and extensive population ageing will occur in all world regions. The most extreme levels will be reached in the Pacific OECD (mostly Japan), where half of the population is likely to be age 60 and above by the end of the century, with the 80 per cent uncertainty interval

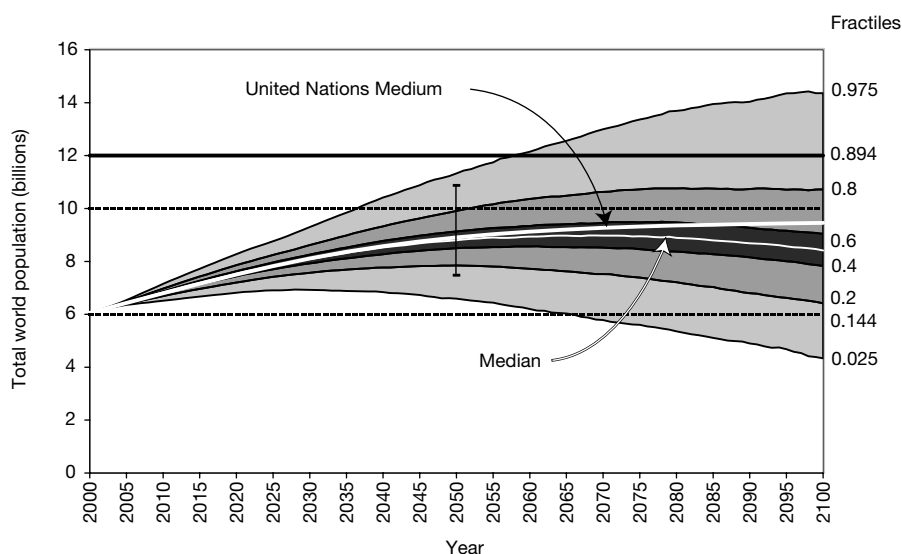


Figure 2 Forecasted distributions of world population sizes (fractiles). For comparison, the United Nations medium scenario (white line), and 95 per cent interval as given by the NRC¹¹ on the basis of an ex post error analysis (vertical line in 2050) are also given.

reaching from 35 to 61 per cent. Even sub-Saharan Africa in 100 years is likely to be more aged than Europe today. The trend of our median proportion over age 60 is almost identical to that of the UN long-range projections² up to 2050, but shows significantly stronger ageing thereafter. This confirms recent criticism that conventional projections tend to underestimate ageing^{6,7}. The extent of and regional differences in the speed of population ageing—the inevitable consequence of population stabilization and decline—will pose major social and economic challenges.

However, population numbers are only one aspect of human impact, and in some of the world's most vulnerable regions, significant population growth is still to be expected. Nevertheless, the prospect of an end to world population growth is welcome news for efforts towards sustainable development. □

Methods

The method of probabilistic population projection that was applied here (see Box 1) is a further development of our earlier approach^{1,8,9} that allows short-term fluctuations in the vital rates^{6,10} and refers to the ex post error analysis of past projections¹¹. We produced a set of 2000 simulations by single years of age for 13 world regions¹² starting in 2000.

Information on baseline conditions has been derived from the United Nations² and US Census Bureau⁴ estimates, and the sensitivity of our results to possible baseline errors is discussed in the Supplementary Information. World population sizes at five-year intervals for all 2000 simulations are also listed there.

The substantive assumptions about future trends in the three components of fertility, mortality and migration, and their associated uncertainty ranges are based on revisions and updates of our earlier work¹² and the extensive analyses summarized in the recent US National Research Council (NRC) report¹¹.

Fertility

The key determinant of the timing of the peak in population size is the assumed speed of fertility decline in the parts of the world that still have higher fertility. On this issue there is

Table 1 Forecasted population sizes and proportions over age 60

Year	Median world and regional population sizes (millions)					Proportion of population over age 60		
	2000	2025	2050	2075	2100	2000	2050	2100
World total	6,055	7,827 (7,219–8,459)	8,797 (7,347–10,443)	8,951 (6,636–11,652)	8,414 (5,577–12,123)	0.10	0.22 (0.18–0.27)	0.34 (0.25–0.44)
North Africa	173	257 (228–285)	311 (249–378)	336 (238–443)	333 (215–484)	0.06	0.19 (0.15–0.25)	0.32 (0.23–0.44)
Sub-Saharan Africa	611	976 (856–1,100)	1,319 (1,010–1,701)	1,522 (1,021–2,194)	1,500 (878–2,450)	0.05	0.07 (0.05–0.09)	0.20 (0.14–0.27)
North America	314	379 (351–410)	422 (358–498)	441 (343–565)	454 (313–631)	0.16	0.30 (0.23–0.37)	0.40 (0.28–0.52)
Latin America	515	709 (643–775)	840 (679–1,005)	904 (647–1,202)	934 (585–1,383)	0.08	0.22 (0.17–0.28)	0.33 (0.23–0.45)
Central Asia	56	81 (73–90)	100 (80–121)	107 (76–145)	106 (66–159)	0.08	0.20 (0.15–0.25)	0.34 (0.24–0.46)
Middle East	172	285 (252–318)	368 (301–445)	413 (296–544)	413 (259–597)	0.06	0.18 (0.14–0.23)	0.35 (0.24–0.47)
South Asia	1,367	1,940 (1,735–2,154)	2,249 (1,795–2,776)	2,242 (1,528–3,085)	1,958 (1,186–3,035)	0.07	0.18 (0.14–0.24)	0.35 (0.25–0.48)
China region	1,408	1,608 (1,494–1,714)	1,580 (1,305–1,849)	1,422 (1,003–1,884)	1,250 (765–1,870)	0.10	0.30 (0.24–0.37)	0.39 (0.27–0.53)
Pacific Asia	476	625 (569–682)	702 (575–842)	702 (509–937)	654 (410–949)	0.08	0.23 (0.18–0.29)	0.36 (0.26–0.49)
Pacific OECD	150	155 (144–165)	148 (125–174)	135 (100–175)	123 (79–173)	0.22	0.39 (0.32–0.47)	0.49 (0.35–0.61)
Western Europe	456	478 (445–508)	470 (399–549)	433 (321–562)	392 (257–568)	0.20	0.35 (0.29–0.43)	0.45 (0.32–0.58)
Eastern Europe	121	117 (109–125)	104 (86–124)	87 (61–118)	74 (44–115)	0.18	0.38 (0.30–0.46)	0.42 (0.28–0.57)
European part of the former USSR	236	218 (203–234)	187 (154–225)	159 (110–216)	141 (85–218)	0.19	0.35 (0.27–0.44)	0.36 (0.23–0.50)

80 per cent prediction intervals are shown in parentheses.

Box 1

Different approaches to probabilistic population projections

The cohort component method of projection is taken as a standard; thus, the differences between alternative approaches discussed in this box refer only to the modelling of future fertility, mortality and migration rates. Here we can distinguish between the specific process chosen for representing the time series of rates, and the basis for the specific assumptions made about the future range of uncertainty.

Model

In the literature there are essentially two methods of specifying the series of vital rates: (1) processes with annual fluctuations^{10,13,14,15}; and (2) piece-wise linear scenarios^{1,8,16}. Whereas method (2) has the advantage of conforming to the current practice of scenario definition in statistical offices around the world (including the UN)², method (1) can produce realistic annual fluctuations given that the possible levels are bounded. We have chosen the following moving-average model with annual fluctuations, in order to avoid the argument that our model underestimates variance¹⁰.

Let v be a vital rate to be forecasted for periods 1 through T and v_t the forecasted value at time t . $v_t = \bar{v}_t + \epsilon_t$, where the mean of v_t , $\bar{v}_t$, and its standard deviation at time t , $\sigma(\epsilon_t)$, are determined according to the assumptions in the text. Let $\{x_{2-n}, \dots, x_T\}$ be the values of $T+n-1$ independent draws from a standard normal distribution and n be the number of periods in the moving average. Then $\epsilon_t = [\sigma(\epsilon_t)/\sqrt{n}] \cdot \sum_{i=t-n+1}^t x_i$. A more detailed description of the model is given in the Supplementary Information.

Assumptions

The literature suggests three approaches for deriving assumptions about the future range of uncertainty of the components: (1) to compute a measure of the future error from the ex post analysis of

past projections^{17,18,19}; (2) to apply time series models^{10,13}; and (3) to have well informed experts make assumptions based on explicitly stated substantive arguments⁹. These three approaches are not mutually exclusive, and approaches (1) and (2) also include expert judgement.

Here we use a synthesis of the three approaches. Our process specification uses a time series model. We have explicitly considered existing national-level parameter estimates^{13,14} given that, at the level of world regions, empirical estimation is impossible owing to lack of data. The ex post analysis of past errors enters our study in two ways: the substantive assumptions made on fertility and mortality changes are informed by the analysis of past errors in those components^{11,18}, and our results at the regional level have been compared to the results of an ex post error analysis of global UN projections documented in the NRC report. Because we preferred to err on the side of higher variance (that is, lower probability of population growth ending this century), we followed the general rule of producing intervals that are at least as large as those in the NRC report at the level of major world regions¹¹. Combining this with argument-based expert judgement¹², we saw substantive reasons for assuming a larger uncertainty in many regions as a result of new factors such as HIV/AIDS, the new situation in the former USSR and the indeterminacy of long-range post-transitional fertility levels that will affect an increasing number of countries.

The 95 per cent interval resulting from the NRC ex post error analysis is inserted in Fig. 2 as a vertical line in 2050 (the latest year given in the NRC report). It corresponds to roughly 80 per cent of our distribution, which clearly indicates that our method produces a broader uncertainty range than the ex post error approach.

a broad consensus that fertility transitions are likely to be completed in the next few decades¹¹. For the eventual size of the population and the question of whether or not world population will begin a decline by the end of this century the key variable is the assumed level of post-transitional fertility. The thorough review of the literature on that subject by the NRC states that "fertility in countries that have not completed transition should eventually reach levels similar to those now observed in low fertility countries" (page 106 in ref. 11). Our fertility assumptions are consistent with this view.

The trends in the means of the regional fertility levels have been defined for the periods 2025–29 and 2080–84 with interpolations in between. The total fertility rates assumed for 2025–29 are similar to those chosen by the United Nations², but for 2080–84 they are assumed to range between 1.5 and 2.0, with lower levels for regions with higher population density in 2030. The variances in the total fertility rates are assumed to depend on the level of fertility. If the total fertility rate is above 3.0 there is an 80 per cent chance that fertility would be within one child of the mean. When it is below 2.0, the same probability is attached to a range within one half a child of the mean. Between the two total fertility rate levels, the variance is interpolated.

Life expectancy

We assume that life expectancy at birth will rise in all regions, except in sub-Saharan Africa, where HIV/AIDS will lower life expectancies during the early part of the century. In general, we assume that life expectancy increases by two years per decade with an 80 per cent probability that the increase is between zero and four years; but there are a number of exceptions to this rule based on specific regional conditions. These assumptions reflect the very large uncertainty that exists regarding future mortality conditions. On one hand, significant biomedical breakthroughs are likely to be made; on the other, AIDS could still become a major issue outside Africa, and new and unexpected threats to human life can emerge.

Autocorrelations

Migration is treated as a random vector on the basis of recent interregional migration patterns. The autocorrelation chosen for all components is based on a 31-year moving-average process that seemed the most plausible after we had experimented with 21-, 31-, and 41-year moving averages (see Supplementary Information), and is close to existing national level figures¹⁰. We assumed an interregional correlation of 0.7 for fertility and 0.9 for mortality deviations with no correlation between fertility and mortality deviations from the assumed trend, and perfect correlation between male and female life expectancy. These choices followed extensive sensitivity analyses as documented in the Supplementary Information. The main rationale behind our choice is that under post-transition conditions, correlations between deviations from assumed fertility and mortality trends are unlikely to be large, while globalization of communication is likely to bring correlated fluctuations of rates among world regions. Mortality correlations will be higher than fertility correlations, owing to the faster communication of medical technology and the faster spread of new health hazards. The sensitivity analysis documented in the

Supplementary Information shows that our main conclusion, that there is around an 85 per cent chance that a peak in world population size will occur in this century, is quite robust to plausible changes in those correlations.

Received 30 April; accepted 22 May 2001.

1. Lutz, W., Sanderson, W. & Scherbov, S. Doubling of world population unlikely. *Nature* **387**, 803–805 (1997).
2. United Nations. *Long-Range World Population Projections: Based on the 1998 Revision* (United Nations, New York, ESA/P/WP.153, 1999).
3. O'Neill, B. C., Scherbov, S. & Lutz, W. The long-term effect of the timing of fertility decline on population size. *Pop. Dev. Rev.* **25**, 749–756 (1999).
4. US Census Bureau. *International Database 10 May 2000 edition* (<http://www.census.gov/ipc/www/worldpop.html>) (US Census Bureau, 2000).
5. World Bank. *World Development Indicators* (CD-ROM) (World Bank, Washington DC, 2000).
6. Tuljapourkar, S., Li, N. & Boe, C. A universal pattern of mortality decline in the G7 countries. *Nature* **405**, 789–792 (2000).
7. Vaupel, J. W. & Lundström, H. in *The Future Population of the World. What Can We Assume Today?* (ed. Lutz, W.) rev. edn, 278–295 (Earthscan, London, 1996).
8. Lutz, W., Sanderson, W. & Scherbov, S. in *The Future Population of the World: What Can We Assume Today?* (ed. Lutz, W.) rev. edn, 397–385 (Earthscan, London, 1996).
9. Lutz, W., Sanderson, W. C. & Scherbov, S. in *Frontiers of Population Forecasting* (eds Lutz, W., Vaupel, J. W. & Ahlburg, D. A.; *Pop. Dev. Rev. Suppl.* **24**, 1998) 139–155 (Population Council, New York, 1999).
10. Lee, R. D. in *Frontiers of Population Forecasting* (eds Lutz, W., Vaupel, J. W. & Ahlburg, D. A.; *Pop. Dev. Rev. Suppl.* **24**, 1998) 156–190 (Population Council, New York, 1999).
11. Bongaarts, J. & Bulatao, R. A. (eds) *Beyond Six Billion. Forecasting the World's Population* (National Academy Press, Washington DC, 2000).
12. Lutz, W. (ed.) *The Future Population of the World. What Can We Assume Today?* rev. edn (Earthscan, London, 1996).
13. Lee, R. & Tuljapourkar, S. Stochastic population projections for the United States: Beyond high, medium and low. *J. Am. Stat. Assoc.* **89**, 1175–1189 (1994).
14. Lee, R. D. & Carter, L. Modeling and forecasting the time series of U.S. mortality. *J. Am. Stat. Assoc.* **87**, 659–671 (1992).
15. Alho, J. M. Stochastic methods in population forecasting. *Int. J. Forecast.* **6**, 521–530 (1990).
16. Lutz, W. & Scherbov, S. An expert-based framework for probabilistic national population projections: The example of Austria. *Eur. J. Pop.* **14**, 1–17 (1998).
17. Alho, J. M. Scenarios, uncertainty and conditional forecasts of the world population. *J. R. Stat. Soc. Ser. A* **160**, 71–85 (1997).
18. Keilman, N. in *Frontiers of Population Forecasting* (eds Lutz, W., Vaupel, J. W. & Ahlburg, D. A.; *Pop. Dev. Rev. Suppl.* **24**, 1998) 15–41 (Population Council, New York, 1999).
19. Stoto, M. The accuracy of population projections. *J. Am. Stat. Assoc.* **78**, 13–20 (1983).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Correspondence and requests for materials should be addressed to W.L. (e-mail: lutz@iiasa.ac.at).

Retrospective and prospective coding for predicted reward in the sensory thalamus

Yutaka Komura^{*,†}, Ryoi Tamura^{*}, Teruko Uwano^{*}, Hisao Nishijo^{*}, Kimitaka Kaga[†] & Taketoshi Ono^{*}

^{*} Department of Physiology, Faculty of Medicine, Toyama Medical and Pharmaceutical University, Toyama 930-0194, Japan

[†] Department of Otorhinolaryngology, University of Tokyo School of Medicine, Tokyo 113-0033, Japan

Reward is important for shaping goal-directed behaviour^{1–4}. After stimulus–reward associative learning, an organism can assess the motivational value of the incoming stimuli on the basis of past experience (retrospective processing), and predict forthcoming rewarding events (prospective processing)^{1–5}. The traditional role of the sensory thalamus is to relay current sensory information to cortex. Here we find that non-primary thalamic neurons respond to reward-related events in two ways. The early, phasic responses occurred shortly after the onset of the stimuli and depended on the sensory modality. Their magnitudes resisted extinction and correlated with the learning experience. The late responses

gradually increased during the cue and delay periods, and peaked just before delivery of the reward. These responses were independent of sensory modality and were modulated by the value and timing of the reward. These observations provide new evidence that single thalamic neurons can code for the acquired significance of sensory stimuli in the early responses (retrospective coding) and predict upcoming reward value in the late responses (prospective coding).

The motivational control of goal-directed action requires the extraction of reward information from various environmental stimuli^{1–4}. On the basis of both clinical and experimental evidence, motivation and reward are associated with the mesolimbic system, striatum and frontal cortex^{2–4,6–15}. In these structures, no specialized peripheral receptors exist for rewards². How and where does the external sensory information meet the internal motivation in the brain? The posterior thalamus, projecting to the amygdala and temporal cortex^{16,17}, seems to have a potential role in aversive information processing^{18–20}. To explore thalamic functions in processing rewarding information, we recorded the activity of single neurons from the posterior thalamic region of well trained rats performing a delayed stimulus–reward association task. The task required discrimination between cue stimuli that did and did not predict rewards. Auditory or visual cue stimuli were presented for 2 s and, after a 1-s delay, a spout automatically protruded close to the mouth of the rat. If the rat licked the spout within 2 s after the delay period, the rat could obtain sucrose solution or intracranial

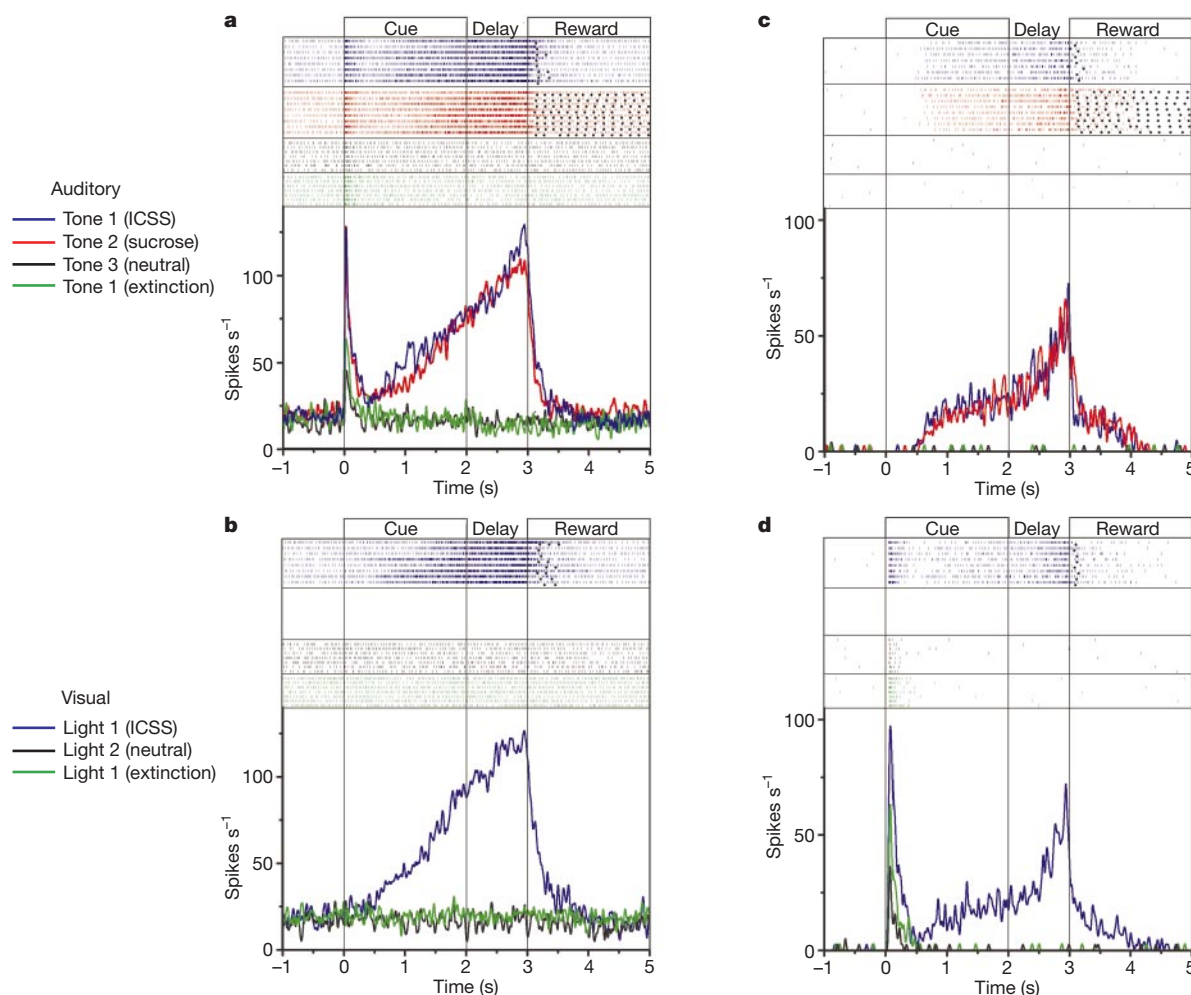


Figure 1 Rastergrams and spike-density functions (SDFs) of the neuronal activity to each cue stimulus (top panels, auditory cues; bottom panels, visual cues). Each dot below a raster line indicates one lick. **a, b**, Dual response to an auditory cue of a representative

neuron in the posterior intralaminar nucleus. **c, d**, The similar response of a representative neuron in the lateral posterior nucleus, except that the cue was a visual stimulus.

self-stimulation (ICSS) in the medial forebrain bundle as a reward.

Figure 1 shows two representative examples of responsive neurons in the posterior thalamic region. These neurons exhibited two temporally separated response components to auditory (Fig. 1a, b) or visual (Fig. 1c, d) stimuli that predicted a reward: the early component at the onset of stimulus presentation and the late component before the reward delivery. The early component was suppressed and the late component disappeared if the stimulus did not predict a reward. We isolated and tested a total of 593 neurons in the posterior thalamic region. Of these cells, 377 responded to one or more of the conditioned stimuli in this task. Of these responsive neurons, 185 showed responses with the two response components shown in Fig. 1.

We next looked at the type of information that is conveyed in the early and late components of the responses. First, we investigated the effects of a change in value of the reward on the activity of the neurons with two response components (Fig. 2a). The same cue stimulus was associated with five different rewards: water (50 μ l); two different amounts of sucrose solution (50 and 100 μ l); and two different intensities of ICSS (70 and 140 μ A). Reward value was assumed to increase in the order²¹: water, <50 μ l sucrose solution and <100 μ l sucrose solution for the natural rewards; and 70 μ A ICSS and <140 μ A ICSS for the artificial rewards. For the neuron in Fig. 2a, as reward value increased, the magnitude of the late response component also increased (one-way analysis of variance (ANOVA)

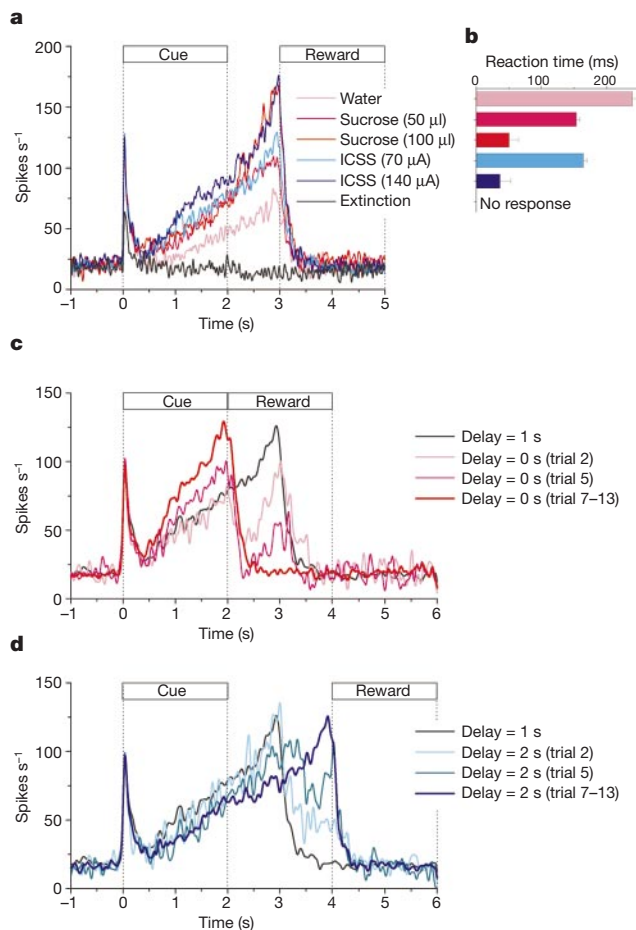


Figure 2 Effects of reward value and timing on the activity of the neuron with two response components, shown in Fig. 1a. The cue was tone 1. **a**, SDFs generated by different reward values. **b**, Histograms of reaction time (the interval between protrusion of the spout and the onset of licking). **c**, SDFs from the 2nd and 5th trials after changing the reward delay from 1 to 0 s. Trial 7–13 is the average SDF from trials 7–13 after changing the delay. **d**, SDFs from the 2nd and 5th trials after changing the delay from 1 to 2 s. Trial 7–13 is the average SDF from trials 7–13 after changing the delay.

$F_{4,33} = 45.2$, $P < 0.001$, with *post hoc* Bonferroni test, $P < 0.001$), but the magnitude of the early response did not. Moreover, as the value of the reward increased, the latency to first lick, that is reaction time, shortened (one-way ANOVA, $F_{4,33} = 52.3$, $P < 0.001$, with *post hoc* Bonferroni test, $P < 0.001$) (Fig. 2b). We tested a total of 56 neurons (35 auditory and 21 visual) with two response components in the same way. In 43 neurons (28 auditory and 15 visual), the magnitude of the late response was modulated significantly by reward value, and the reward value was negatively correlated with the reaction time of the initial lick.

Next, we examined the effects of reward delay time on the activity of these neurons. The delay between the end of the cue and the onset of reward was set as 0, 1 or 2 s. If the reward delay was reduced from 1 to 0 s (Fig. 2c) or increased from 1 to 2 s (Fig. 2d), a peak of the late response shifted to the left from approximately 3 to 2 s or the right from approximately 3 to 4 s, respectively, after the onset of the cue. During the first seven trials after the delay period was altered from 1 to 0 or 2 s, the late component displayed two peaks, which were not necessarily accompanied by licking. In contrast to the late component, reward timing did not modulate the peak of the early component ($F_{2,21} = 0.33$, not significant). We found similar adaptations to reward timing in the peak of the late responses of all 41 neurons tested (25 auditory and 16 visual).

Third, we assessed the plasticity of the cells during extinction of learning and relearning of stimulus–reward association (Fig. 3). After eight pre-extinction trials (block 1), the rat continued to lick the spout during the first eight extinction trials (block 2), and then ceased licking the spout in the following trials (blocks 3–9). During these extinction trials, the magnitude of the early responses decreased slowly to reach a stable level by block 7, which was larger than the response to the neutral stimuli. The magnitude of the late response, on the other hand, decreased rapidly and did not differ from spontaneous activity by block 3. For the next six blocks (4–9), the neural activity did not change significantly. In the next

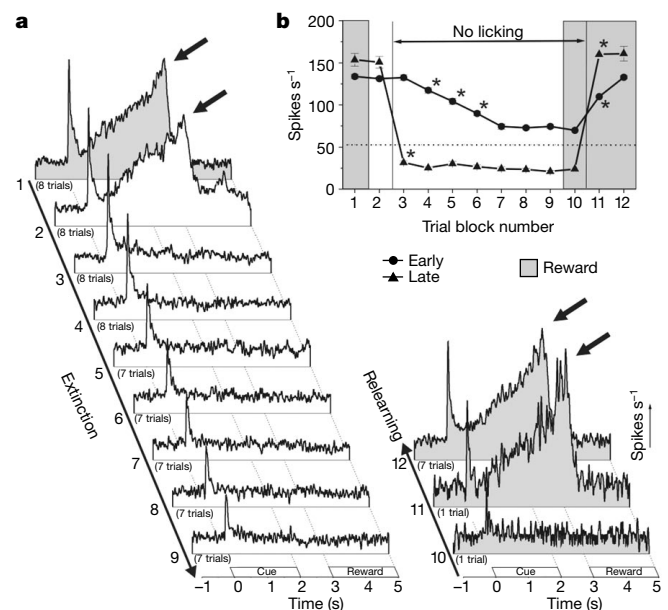


Figure 3 The plasticity of responses to tone 1 of the neuron shown in Fig. 1a during extinction and relearning. The shaded areas indicate trials with reward delivery. **a**, SDFs across blocks of trials. Each block consists of 7 or 8 trials, except for blocks 10 and 11 (single trial). Arrows indicate trials during which the rat licked the spout after the delay period. Scale bar, 50 spikes s⁻¹. **b**, Change in the magnitude of the early and late responses. Each asterisk shows that the response magnitudes were significantly different than those of each previous block (one-way ANOVA $F_{11,64} = 64.9$, $P < 0.001$, with *post hoc* Bonferroni test, all $P < 0.001$). The dotted line indicates the firing level of the early response to a neutral auditory stimulus.

trial (block 10), an ICSS reward was delivered after the cue, regardless of the rat's behavioural response, to reinstitute the association between the stimulus and the reward. The rat began to lick the spout again (block 11), and the magnitude of both the early and late responses recovered rapidly to pre-extinction levels. The magnitude of the late response correlated well with the behavioural responses: both rapidly decreased and increased during extinction and relearning, respectively. On the other hand, the magnitude of the early component slowly decreased during extinction, but rapidly recovered during relearning. Extinction does not erase the memory of conditioning²², as the behavioural response returns readily with relearning.

The asymmetrical change in the early response during extinction and relearning indicated that this component was influenced by the memory of a previously acquired stimulus–reward association. Even after behavioural extinction, the early response to the conditioned stimuli ($n = 185$ neurons) was significantly greater than that to the neutral stimuli (paired t -test, $P < 0.001$). Furthermore the responses to the highest and the lowest frequency cue tones, which were associated with rewards, were greater than to the intermediate,

non-rewarded tones (one-way ANOVA, $F_{2,405} = 988.8$, $P < 0.001$, with *post hoc* Bonferroni test, $P < 0.001$) and equivalent to one another ($P = 0.11$). The results are different from previous studies^{23,24} on the sensory properties of the posterior thalamus of naive rats that demonstrated that neurons are more sensitive to higher frequency tones within the same frequency range that was used in our study. Therefore, differences of the early component documented here probably result from the experiences of reward contingencies rather than the physical properties of the stimuli.

Finally we investigated the functional anatomy of these responses. Thalamocortical connections are categorized into two parallel primary and non-primary pathways, and the posterior thalamic region contributes to both of these (Fig. 4a)^{16,17,25,26}. All of the 185 neurons with two response components (Fig. 1) were located in the non-primary thalamic nuclei. Of these, the neurons recorded in the dorsal or medial part of the medial geniculate nucleus and the posterior intralaminar nucleus coded predominantly auditory stimuli in the early phase. The neurons in the lateral posterior nucleus and supragenicular nucleus coded visual stimuli. In contrast, the cells in the primary thalamic region showed a clear modality specificity without reward dependency: the ventral medial geniculate nucleus neurons and the dorsal lateral geniculate nucleus responded only to the auditory and visual stimuli, respectively, and only during the cue presentation. The magnitude of the responses of these cells did not change after the manipulation of reward contingency. Figure 4b summarizes the relationships of sensory specificity and reward dependency for the responsive neurons in the sample.

How does the dual temporal coding in the posterior thalamus work together with processing reward information in other brain areas^{2–4,6–15}? The functional distribution shown in Fig. 4b is consistent with the fact that only the non-primary thalamus sends efferents directly to components of the reward circuit, such as the amygdala, striatum and perirhinal cortex (Fig. 4a)^{16,17}. Because the magnitude of the early components varied with the conditioning experience (Figs 1 and 3), the posterior thalamus might serve as a filter for the forebrain that favours behaviourally significant stimuli over the others by amplifying response magnitude. This hypothesis also corresponds with a theoretical model and with human studies that stipulate the thalamus, especially the primate pulvinar (homologue of the rat lateral posterior nucleus²⁶), selectively enhances attended sensory signals^{20,27}. On the other hand, the climbing activity that characterizes the late responses may code prospectively for the anticipated reward. These responses probably arise from interaction with the other brain areas contributing to reward expectation^{2–4,8,9,11,12,14,15}, via the perirhinal cortex. This is because the perirhinal cortex seems to be important in attaching motivational significance to stimuli²⁸, and among the constituents of the presumed reward circuits, only the perirhinal cortex is known to send the direct inputs to the non-primary thalamus¹⁶. In this way, the posterior thalamus could communicate with the reward circuits, through the dual temporal coding for mnemonic and anticipatory process.

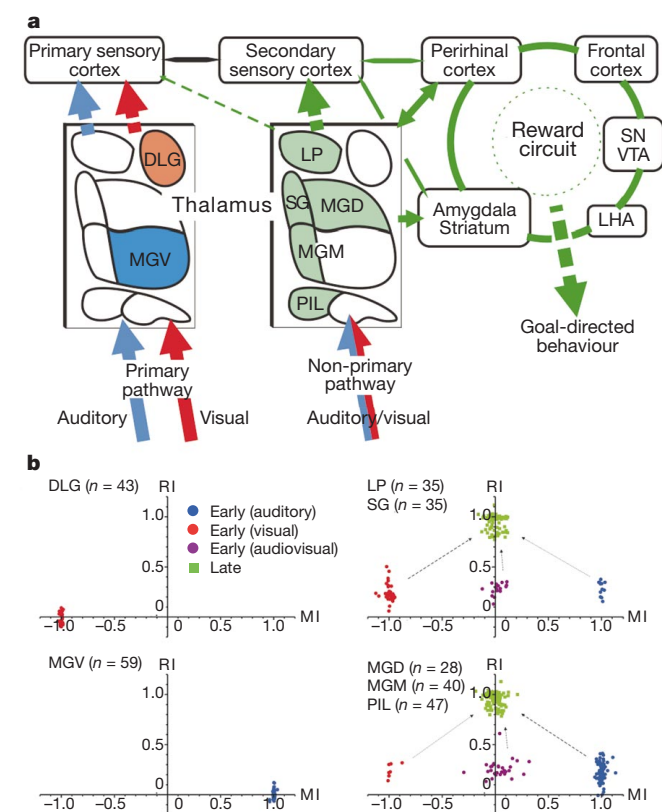


Figure 4 Functional topography. **a**, Parallel pathways through the auditory and visual thalamic nuclei (recording sites). In the auditory system, the primary pathway ascends from the ventral part of the medial geniculate nucleus (MGV) to the primary auditory cortex. The non-primary pathway ascends from the dorsal or medial part of the medial geniculate nucleus (MGD and MGM, respectively), supragenicular nucleus (SG) and posterior intralaminar nucleus (PIL) to the secondary auditory cortex. In the visual system, the primary pathway is from the dorsal lateral geniculate nucleus (DLG) to the primary visual cortex. The extrageniculate or non-primary pathway is from the lateral posterior nucleus (LP) to the secondary visual cortex. SN, substantia nigra; VTA, ventral tegmental area; LHA, lateral hypothalamic area. **b**, The effects of the sensory specificity and reward association in each nucleus. The indices (MI and RI) of the neuron shown in Fig. 1a and b were 0.98 and 0.22, respectively, for the early component and -0.06 and 1.00 , respectively, for the late component. MI and RI values of the neuron shown in Fig. 1c and d were -1.01 and 0.23 , respectively, for the early component and -0.01 and 1.00 , respectively, for the late component.

Methods

Animal care

The experimental procedures were similar to those used previously^{9,11}, and the experiments were conducted in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals, and with Guideline for the Care and Use of Laboratory Animals in Toyama Medical and Pharmaceutical University.

Animal preparation

Twelve male Wistar rats (270–330 g) were anaesthetized (sodium pentobarbital 40 mg per kg, intraperitoneal injection) and had a cranioplastic cap attached to the skull. This made it possible to fix the animal's head painlessly in the stereotaxic device. At the same time, two concentric bipolar electrodes for ICSS were implanted in the medial forebrain bundle (anterior, -3.8 from bregma; lateral, ± 1.4 mm; ventral, 8.4 mm or anterior, -5.2 mm; lateral, ± 0.8 mm; ventral, 8.2 mm) according to the atlas of Paxinos and Watson²⁹. A hole

(3.0–5.0 mm diameter) for chronic recording was drilled through the cranioplastic cap and the underlying skull.

Behavioural task

The rats were required to lick a spout to obtain rewards of 0.3 M sucrose solution or ICSS (0.5-s train of 100-Hz, 0.3-ms capacitor-coupled negative square wave pulses). We monitored licking behaviour with a photoelectric sensor. The threshold level for ICSS to maintain the licking behaviour in the operant task was determined in behavioural tests before recording. In this study, the thresholds ranged from 65 to 110 μ A. The licking frequency depended on the method of reward delivery, that is, low for ICSS and high for a liquid reward. This was probably because ICSS was delivered only once at the first lick even if the rat licked several times, whereas liquid was delivered continuously during a 2-s reward period, although the property of the reward, natural or artificial, could also be involved in this difference. In the cued operant task, a 1,200 Hz tone (tone 1) and a white light on the left side (light 1) signalled availability of the ICSS, and a 4,300 Hz tone (tone 2) signalled availability of sucrose solution. A 2,800 Hz tone (tone 3) and a white light on the right side (light 2) were neutral, that is, not associated with reward. A speaker positioned 10 cm anterior to the rats delivered tones that were 85 dB at each ear and two white light lamps, 45° lateral and 7 cm from each eye, delivered the visual stimuli.

Electrophysiology

Before the unit recording session, we searched for auditory neural responses by clapping and, thus, localized the margins of the medial geniculate body. On the basis of the coordinates derived from this localizing procedure, a glass-insulated tungsten micro-electrode ($Z = 1.0$ – 1.5 M Ω at 1,000 Hz) was inserted stereotactically stepwise by a micromanipulator into various parts of the posterior thalamus. Extracellular discharges of single neurons were recorded using conventional recording procedure.

Histology

At the end of recording session, several small electrolytic lesions (20 μ A for 20 s) were made stereotactically around the recording sites with a glass-insulated tungsten micro-electrode. The brains were sectioned coronally (30 μ m) and stained with cresyl violet. All marking and stimulation sites were then verified microscopically. The positions of the recorded neurons were mapped onto the appropriate tissue sections, using stereotaxic coordinates, and the sections were compared with the atlas of Paxinos and Watson²⁹.

Data analysis

Spike trains were smoothed by convolution with a gaussian kernel ($\sigma = 20$ ms) to obtain spike-density functions (SDFs)³⁰. Spontaneous activity was defined as the mean discharge rate during the 1,000 ms preceding cue onset. The early response was defined as the number of spike counts in the first 100 ms after cue onset. The late response was defined as the number of spike counts in the 500 ms period from 2,500 to 3,000 ms after cue onset. A neuron was classified as responsive to any cue stimuli if the difference in neural activity in three periods (spontaneous activity, early response and late response) was significant (one-way ANOVA, $P < 0.01$). A modality coupling index (MI) and a reward coupling index (RI) were computed for each response component: $MI = (F_{\text{aud(ICSS)}} + F_{\text{aud(ext)}} - F_{\text{vis(ICSS)}} - F_{\text{vis(ext)}})/F_{\text{sum}}$; $RI = (F_{\text{aud(ICSS)}} - F_{\text{aud(ext)}} + F_{\text{vis(ICSS)}} - F_{\text{vis(ext)}})/F_{\text{sum}}$ where $F_{\text{sum}} = F_{\text{aud(ICSS)}} + F_{\text{aud(ext)}} + F_{\text{vis(ICSS)}} + F_{\text{vis(ext)}}$. $F_{\text{aud(ICSS)}}$ or $F_{\text{vis(ICSS)}}$ is the average firing rate to auditory stimulus or visual stimulus, respectively, associated with reward minus spontaneous firing rate in each neuron. $F_{\text{aud(ext)}}$ or $F_{\text{vis(ext)}}$ is the average firing rate to auditory stimulus or visual stimulus, respectively, in the extinction test minus spontaneous firing rate in each neuron. MI represents sensory modality selectivity of the neural responses and RI represents reward dependency. For example, if the value of MI was 1.0 or -1.0 , this indicates a purely auditory-specific response or a visual one, respectively. An RI value of 1.0 or 0 indicates that a neural response is totally dependent on or independent of reward contingency, respectively.

In principle, memory and anticipation are mutually dependent. However, the temporal response patterns (Fig. 1) and the effects of manipulating reward parameters (Figs 2 and 3) suggested that the early and late components conveyed distinguishable information. Therefore we designated the early component as related to memory (retrospective coding) and the late one as related to anticipation (prospective coding).

Received 26 January; accepted 13 June 2001.

- Dickinson, A. & Balleine, B. Motivational control of goal-directed action. *Anim. Learn. Behav.* **22**, 1–18 (1994).
- Schultz, W. Multiple reward signals in the brain. *Nature Rev. Neurosci.* **1**, 199–207 (2000).
- Schultz, W., Tremblay, L. & Hollerman, J. R. Reward processing in primate orbitofrontal cortex and basal ganglia. *Cereb. Cortex* **10**, 272–283 (2000).
- Rolls, E. T. *The Brain and Emotion*. (Oxford Univ. Press, New York, 1999).
- Rainer, G., Rao, S. C. & Miller, E. K. Prospective coding for objects in primate prefrontal cortex. *J. Neurosci.* **19**, 5493–5505 (1999).
- Fibiger, H. C. & Phillips, A. G. in *Handbook of Physiology—The Nervous System* Vol. IV (ed. Bloom, F. E.) 647–675 (Williams and Wilkins, Baltimore, 1986).
- Wise, R. A. & Rompre, P. P. Brain dopamine and reward. *Annu. Rev. Psychol.* **40**, 191–225 (1989).
- Mirenzowicz, J. & Schultz, W. Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature* **379**, 449–451 (1996).
- Nakamura, K., Ono, T. & Tamura, R. Central sites involved in the lateral hypothalamus conditioned neural responses to acoustic cues in the rat. *J. Neurophysiol.* **58**, 1123–1148 (1987).
- Robbins, T. W. & Everitt, B. J. Neurobehavioral mechanism of reward and motivation. *Curr. Opin. Neurobiol.* **6**, 228–236 (1996).

- Ono, T., Nishijo, H. & Uwano, T. Amygdala role in conditioned associative learning. *Prog. Neurobiol.* **46**, 401–422 (1995).
- Schoenbaum, G., Chiba, A. A. & Gallagher, M. Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature Neurosci.* **1**, 155–159 (1998).
- Bechara, A., Damasio, H., Tranel, D. & Anderson, S. W. Dissociation of working memory from decision making within the human prefrontal cortex. *J. Neurosci.* **18**, 428–437 (1998).
- Watanabe, M. Reward expectancy in primate prefrontal neurons. *Nature* **382**, 629–632 (1996).
- Kawagoe, R., Takikawa, Y. & Hikosaka, O. Expectation of reward modulates cognitive signals in the basal ganglia. *Nature Neurosci.* **1**, 411–416 (1998).
- LeDoux, J. E., Farb, C. R. & Romanski, L. M. Overlapping projections to the amygdala and striatum from auditory processing areas of the thalamus and cortex. *Neurosci. Lett.* **134**, 139–144 (1991).
- Linke, R., De Lima, A. D., Schwegler, H. & Pare, H.-C. Direct synaptic connections of axons from superior colliculus with identified thalamo-amygdaloid projection neurons in the rat: possible substrates of a subcortical visual pathway to the amygdala. *J. Comp. Neurol.* **403**, 158–170 (1999).
- Weinberger, N. M. Learning-induced changes of auditory receptive fields. *Curr. Opin. Neurobiol.* **3**, 570–577 (1993).
- LeDoux, J. E. Emotion circuits in the brain. *Ann. Rev. Neurosci.* **23**, 155–184 (2000).
- Dolan, R. J. in *The New Cognitive Neurosciences* 2nd edn (ed. Gazzaniga, M. S.) 1115–1131 (MIT Press, Cambridge, Massachusetts, 2000).
- Gallistel, C. R., Leon, M., Waraczynski, M. & Hanau, M. S. Effect of current on the maximum possible reward. *Behav. Neurosci.* **105**, 901–912 (1991).
- Bouton, M. E. Context, ambiguity, and classical conditioning. *Curr. Direct. Psychol. Sci.* **3**, 49–53 (1994).
- Bordi, F. & LeDoux, J. E. Response properties of single units in areas of rat auditory thalamus that project to the amygdala. I. Acoustic discharge patterns and frequency receptive fields. *Exp. Brain Res.* **98**, 261–274 (1994).
- Edeline, J. M., Manunta, Y., Nodal, F. R. & Bajo, V. M. Do auditory responses recorded from awake animals reflect the anatomical parcellation of the auditory thalamus? *Hear. Res.* **131**, 135–152 (1999).
- Winer, J. A. & Morest, D. K. The medial division of the medial geniculate body of the cat: implications for thalamic organization. *J. Neurosci.* **3**, 2629–2651 (1983).
- Sefton, A. J. & Dreher, B. in *The Rat Nervous System* 2nd edn (ed. Paxinos, G.) 833–898 (Academic, San Diego, 1995).
- LaBerge, D. Thalamic and cortical mechanisms of attention suggested by recent positron emission tomographic experiments. *J. Cogn. Neurosci.* **2**, 358–372 (1990).
- Liu, Z., Murray, E. A. & Richmond, B. J. Learning motivational significance of visual cues for reward schedules requires rhinal cortex. *Nature Neurosci.* **3**, 1307–1315 (2000).
- Paxinos, G. & Watson, C. *The Rat Brain in Stereotaxic Coordinates* 4th edn (Academic, San Diego, 1998).
- Richmond, B. J. & Optican, L. M. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. II. Quantification of response waveform. *J. Neurophysiol.* **57**, 147–161 (1987).

Acknowledgements

We thank R. Norgren (invited by Gofu Life Sciences International Fund) for helpful comments on this manuscript, and T. Kitamura for technical assistance. This work was partly supported by Grants-in-Aid for Scientific Research from the Japanese Ministry of Education, Science and Culture.

Correspondence and requests for materials should be addressed to T.O.
(e-mail: onotake@ms.toyama-mpu.ac.jp).

Practising orientation identification improves orientation coding in V1 neurons

Aniek Schoups*, Rufin Vogels*, Ning Qian† & Guy Orban*

*Laboratorium voor Neuro-en Psychofysiologie, K.U. Leuven Medical School, B-3000 Leuven, Belgium

†Center for Neurobiology and Behavior, and Department of Physiology and Cellular Biophysics, Columbia University, New York 10032, USA

The adult brain shows remarkable plasticity, as demonstrated by the improvement in fine sensorial discriminations after intensive practice. The behavioural aspects of such perceptual learning are well documented, especially in the visual system^{1–8}. Specificity for stimulus attributes clearly implicates an early cortical site, where receptive fields retain fine selectivity for these attributes; however, the neuronal correlates of a simple visual discrimination task remained unidentified. Here we report electrophysiological correlates in the primary visual cortex (V1) of monkeys for learning orientation identification. We link the behavioural improvement

in this type of learning to an improved neuronal performance of trained compared to naive neurons. Improved long-term neuronal performance resulted from changes in the characteristics of orientation tuning of individual neurons. More particularly, the slope of the orientation tuning curve that was measured at the trained orientation increased only for the subgroup of trained neurons most likely to code the orientation identified by the monkey. No modifications of the tuning curve were observed for orientations for which the monkey had not been trained. Thus training induces a specific and efficient increase in neuronal sensitivity in V1.

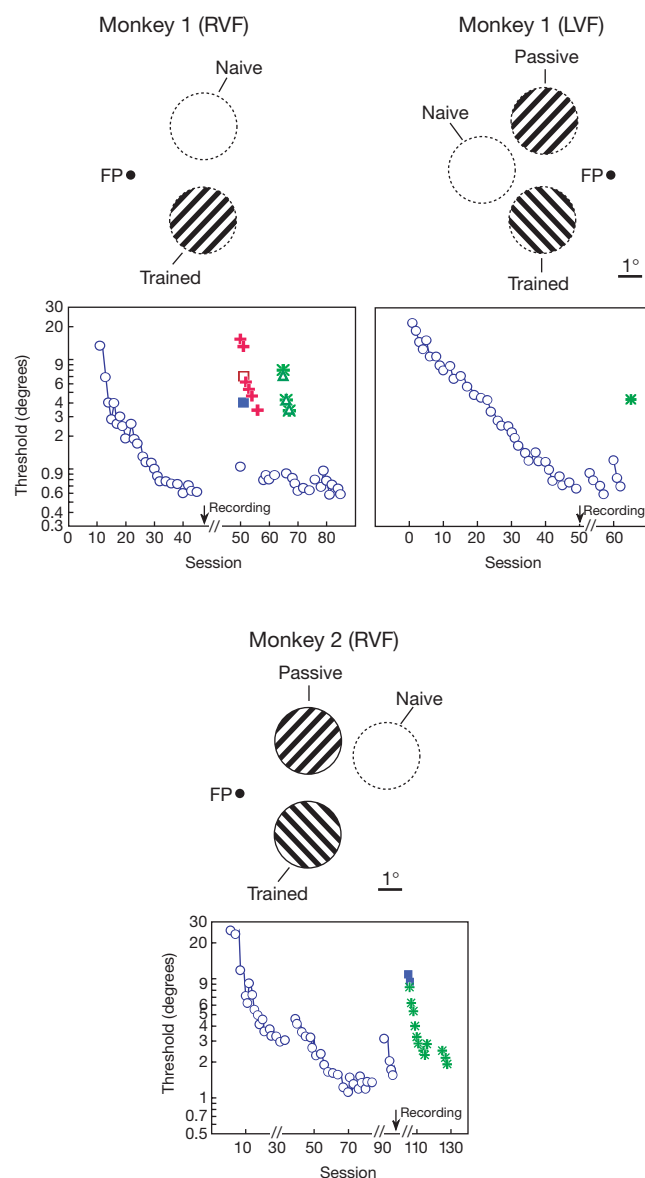


Figure 1 Behavioural performance and specificity for position and orientation. Learning curves for orientation identification for monkey 1 (two hemispheres) and monkey 2 (RVF). Above each learning curve, a schematic representation of the stimulus positions and reference orientation is given for each of the hemispheres studied. Thresholds for the trained orientation and position decreased on consecutive daily sessions (open blue circles). After recording, transfer was tested to the other oblique orientation (filled blue squares) at the same stimulus position, and to other stimulus positions (green, naive position; red, passively stimulated position). Thresholds at these other stimulus positions were similar for the two oblique orientations (green stars and triangles; red plus signs and open square), and were 7–12 times larger than for the trained position and orientation. FP, fixation point.

The psychophysics of early perceptual learning has been well studied; however, researchers are only beginning to understand the neurophysiological correlates of perceptual learning. Primary somatosensory, motor and auditory cortex show learning-dependent changes in their topographical organization^{9–11}. In the visual system, however, early plasticity is demonstrated only by lesion-dependent reorganizations^{12,13}. Improvement in motion discrimination, which transfers to other retinal stimulus positions, is accompanied by short-term improvements in neuronal sensitivity in the higher-order middle temporal and medial superior temporal areas within a single session, but does not extend across sessions^{14,15}.

We trained two monkeys to identify the orientation of a small grating (Fig. 1). The performance of monkeys, like that of human subjects⁷, improved markedly with training, reaching a threshold as low as 0.6–1.2° after several months. The improvement was specific for both stimulus position and orientation. This specificity provided us with an internal control: instead of comparing data between monkeys, we could compare different populations of

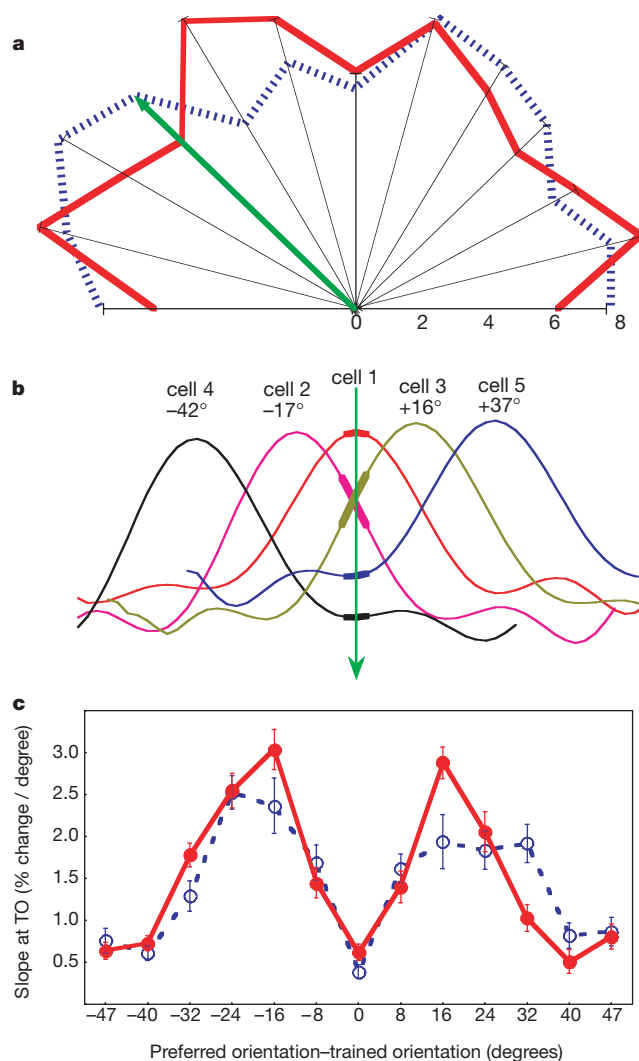


Figure 2 Neuronal responses. **a**, Distribution (%) of preferred orientations in trained (solid red line) and naive neurons (dashed blue line). Data are from all cell layers. The green arrow indicates the trained orientation. **b**, Orientation tuning curves of five sample neurons. **c**, Slope measured at trained orientation (TO) for trained (solid red line) and naive neurons (dashed blue line). Neurons are classified according to the angle between the preferred and trained orientation. Neurons tuned to the trained orientation are shown in the centre, groups with preferred orientation clockwise from trained orientation are to the left; those anticlockwise to the right. Data ($\pm$ s.e.m.) are from layers 2–3 and 5–6 combined.

neurons within each hemisphere. The 'trained' population consisted of neurons with receptive fields at the trained position. A first control group (passive) consisted of cells responding to a second grating presented equally frequently during training, but that was behaviourally ignored by the subject. Second, neurons with receptive fields outside the trained or passive locations were considered naive. The monkeys' performance was verified after the recordings. The thresholds at which they could discriminate untrained orientations at the trained position, or the trained orientation at the two types of control positions, always exceeded the thresholds for the trained orientation at the trained position (Fig. 1).

After training, we recorded from 1,430 single units in V1 while the monkeys were performing a fixation task. The purpose was to reveal changes in the tuning characteristics of trained neurons that could be responsible for the observed improvement in sensory discrimination. Contrary to expectation^{9–11}, we did not observe any shift in preferred orientations in favour of the trained orientation. The optima of neurons were evenly distributed over all orientations both before and after training (Fig. 2a).

It was only when we grouped all cells according to their preferred orientation that a specific change in the response characteristics of the trained neurons became apparent. Figure 2b shows examples of

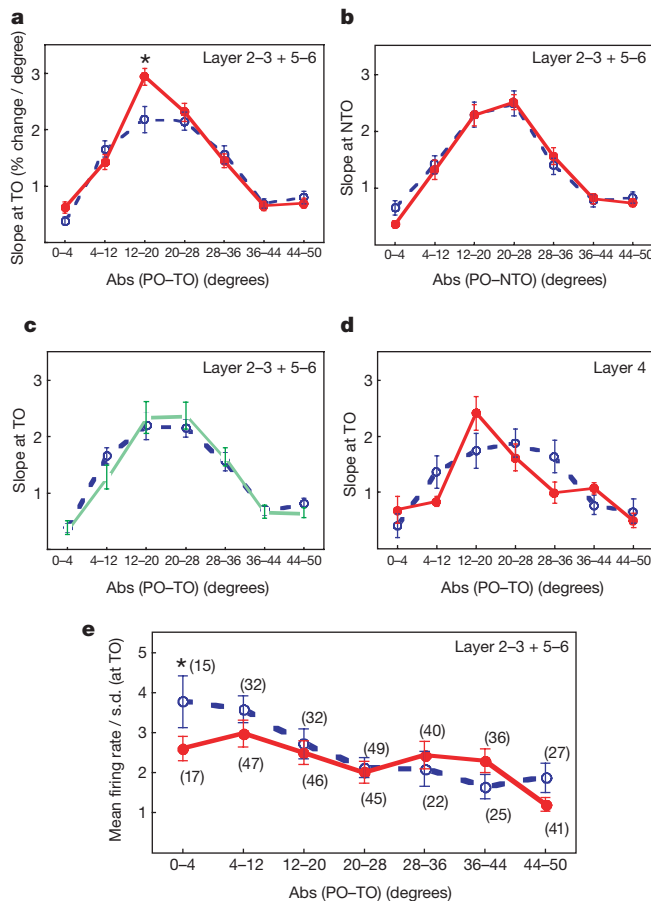


Figure 3 Neuronal responses. **a**, Increase in the slope of the tuning curve measured at the trained orientation (TO) for trained neurons tuned to orientations within 12 to 20° of the trained orientation. **b–d**, This effect was not observed at the NTO (**b**), in passively stimulated neurons (**c**), nor in trained neurons from layer 4 (**d**). **e**, Decrease in the signal-to-noise ratio for trained cells tuned to orientations within 0–4° of the trained orientation. This ratio is given by the mean firing rate (averaged raw data) to the trained orientation, divided by the s.d. of the response distribution. The numbers of neurons for each preferred orientation (PO) group are indicated between brackets. In all panels trained neurons are indicated by a solid red line, naive neurons by dashed blue line, and passively stimulated neurons by dotted green line. *, significantly different.

tuning curves, normalized to their respective maxima, with the preferred orientation indicated above each tuning curve. We measured the slopes of the tuning curves at the trained orientation (indicated by the thickening of the curves in Fig. 2b). The slopes are small, both for neurons that have their optima at or near the trained orientation (cell 1), and for those that do not respond to the trained orientation (for example cells 4 and 5). These neurons provide little information useful for detecting orientation differences. In contrast, the slope at the trained orientation is largest for neurons whose optimum differs by about 20° from the trained orientation (for example cells 2 and 3). The firing rate of these neurons show the greatest change in response to a 1° change in orientation, representing the highest sensitivity to small orientation differences^{16–18}. Figure 2c plots the slope values as a function of angle (preferred orientation to trained orientation (PO-TO)) for cells with preferred orientations lying within 50° of the trained orientation. At the point at which sensitivity is exactly most acute (that is, trained neurons with preferred orientations lying between 12 and 20° of the trained orientation), an increase in the slopes of the tuning curves is observed. The increase in slope was consistent in that it was observed for trained cells with optima lying on either side of the trained orientation (Fig. 2c). In Figure 3a, both sides are combined, and the slope at the trained orientation is shown for the cells grouped according to the absolute value of the angle between the preferred orientation and the trained orientation. A significant main effect of angle (PO-TO) was observed, as well as a significant interaction between training and angle (PO-TO) ($F_{6,492} = 2.88$; $P < 0.009$). *Post hoc* analysis (least significant difference) showed that training has a significant effect only for the group of neurons with optima between 12 and 20° of the trained orientation. This difference was observed in all three hemispheres studied (slopes for trained neurons with preferred orientations between 12 and 20° of the trained orientation: 3.1 ± 0.24 , 3.0 ± 0.25 , 2.7 ± 0.29 ; for naive

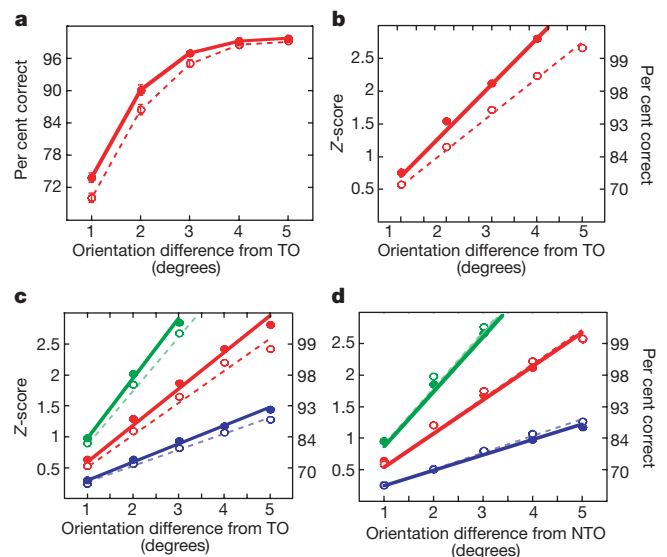


Figure 4 Neuronal performance. **a**, Percentage correct for a set of 20 cells, chosen from all trained or naive neurons tuned to orientations within 45° of the trained orientation (TO), in identifying orientation differences around the trained orientation. **b**, Performance for 20 cells randomly selected from the subset of neurons tuned to 12–20° from the trained orientation. Slope is 27% higher for trained neurons compared with naive neurons. **c**, z-transformed per cent correct as a function of orientation difference for various set sizes (20 cells: red, same data as in **a**; 5 cells: blue; 50 cells: green) randomly selected out of all trained or naive cells. Trained neurons always have a significantly steeper slope than naive neurons. Improvement for 5, 20 and 50 cells amounted to 12.1, 12.5 and 11.6%, respectively. **d**, Same as in **c**, for identifying the untrained oblique orientation. Trained neurons have the same slope as naive neurons, for any set size of neurones selected. For all panels, solid lines indicate trained neurons; dashed lines indicate naive neurons.

neurons: 2.2 ± 0.34 , 2.3 ± 0.39 , 1.9 ± 0.59). As a first control, we performed the same analysis around the untrained oblique orientation (NTO). As described above, the monkeys had much larger thresholds for the NTO, 90° from the trained orientation (Fig. 1, filled blue squares). Consistent with this behaviour, no significant differences were observed between naive and trained neurons for the slope measured at the NTO (Fig. 3b; $F_{6,459} = 0.42$; not significant). Figure 3c further shows that exposure to the stimulus alone was not sufficient to induce a change in the slope at the trained orientation, as passively stimulated cells failed to show this effect. Moreover, the changes in slopes of the tuning curve at the trained orientation occurred in cells in the supra- and infragranular layers, but was not observed in those of input layer 4 (Fig. 3d).

The signal-to-noise ratio at the trained orientation is given by the mean firing rate in response to the trained orientation, divided by the s.d. A decrease in this ratio was observed for trained neurons tuned to orientations near the trained orientation (Fig. 3e; *post hoc* analysis $P < 0.01$). If a monkey were to base its decision on the signals from the most active neurons, performance would deteriorate. The fact that it doesn't indicates once again that the difference signals arising from neurons whose relative activity changes the most sharply are the most relevant for the task^{16–18}. The response variance, normalized to spike counts, remained unchanged after training, and averaged 1.85 and 1.90 spikes per 300 ms for naive and trained cells, respectively. No interaction between preferred orientation group and effect of training was observed ($P = 0.88$).

The increase in the slopes of the orientation tuning curves at the trained orientation was predicted by a recent computational model of orientation learning¹⁹ based on the recurrent models of orientation selectivity^{20,21}. According to the model, reductions in neuronal activity at the trained orientation can lead to specific changes in the tuning curves, exactly as we have observed.

We next asked how relevant the observed changes in the orientation tuning curves to performance in the orientation discrimination task are. We calculated the neuronal discrimination abilities of all naive or trained cells using bayesian analysis^{22,23}. Like any other reconstruction method it estimates the orientation presented from the observed neuronal activity. We used it to calculate the percentage of correct estimates for a given orientation difference. The orientation that has the highest probability of having occurred, given the neuronal response, was compared to the actual presented orientation for two orientations on either side of the trained orientation. The chief advantage of this method is that it performs like an ideal observer of population activity. Figure 4a shows the neuronal discrimination performance of a set of 20 cells randomly selected from all naive or trained cells with preferred orientations within 45° of the trained orientation. The neurometric curve for the trained population is shifted towards smaller orientation differences compared with the naive population. A cumulative normal distribution was fitted to these data points (z -transformation)⁷. After linear regression, only one parameter, the slope, defines the neuronal performance. The significantly greater slope (12% increase; $P < 0.05$, $n = 20$) for trained compared with naive cells in discriminating orientation differences around the trained orientation is shown in Fig. 4c. For increasing population sizes, the threshold decreases at a rate proportional to the square root of the number of cells. This hypothesis holds, provided that the firing of the neurons is not correlated, which was not taken into consideration when performing the bayesian analysis. The improvement in performance of trained versus naive cells, however, was independent of the set size (Fig. 4c). The observed change is quantitatively similar to the 13.6% improvement noted in motion-sensitive neurons of the middle temporal and medial superior temporal areas between two successive blocks of learning trials in a motion-discrimination task¹⁴. The difference between naive and trained neurons was not observed when orientation differences around the NTO were considered (Fig. 4d). Finally, when we consider only those neurons tuned to

orientations between 12 and 20° of the trained orientation, the neuronal improvement that accompanies training is much larger (27% increase in slope; Fig. 4b). Again, the trained group did not differ from naive cells for orientations of 1 to 5° from the NTO (data not shown).

There are many reasons why the difference in neuronal performance between naive and trained cells did not equal the difference in thresholds for trained and untrained positions, observed psychophysically. On one hand, the behavioural improvement is probably overestimated. Improvement in a task like this one encompasses many behavioural aspects, and monkeys in particular are sensitive to any change in stimulus. Similar experiments with human subjects⁷ revealed an improvement in psychophysical performance after learning of only 56% (increase in slope of the z -normalized psychometric curve). On the other hand, V1 may not be the sole locus of the sensorial improvement. Conversely, the neuronal performance could be underestimated. Although we used the most ideal classifier, it still does not take into account response correlations, which can improve discrimination and be affected by learning^{23–25}.

A number of mechanisms have been suggested to explain various forms of perceptual learning. An expansion of the cortical representation has been demonstrated for tasks using features represented in topographical maps such as sound frequency or somatotopic localization^{9–11}. In our study, however, the proportion of cells that preferred the trained orientation remained unchanged. Neuronal noise reduction has also been suggested as an alternative mechanism²⁶. Our data, in accordance with that of others²⁷, do not support this idea. Instead, we found that learning is correlated with specific changes in the tuning curves of specific groups of V1 cells. Only those cells that are relevant for signalling an orientation difference increase their discriminatory ability. Although our data do not preclude a multi-stage involvement, they show a first significant change in primary visual cortex. □

Methods

Psychophysics

We trained two adult male rhesus monkeys (*Macaca mulatta*) in an orientation identification task. Monkey 1 was first trained to identify the right oblique reference orientation positioned in its right visual field (RVF), and after recording, was retrained using the left visual field (LVF) and the left oblique orientation. Monkey 2 was trained in the RVF, using the left oblique orientation. For two out of the three hemispheres, a second stimulus (passive) was placed in the upper visual field of the subject during training. The orientation of the second stimulus was randomly chosen from a narrow range of orientations (0 – 5°) around the oblique reference orientation, orthogonal to the one that had been trained. The monkey learned to ignore this stimulus. The monkeys had a stainless steel head-fixing device mounted to the skull and a scleral search coil implanted²⁸. The stimulus was a sinusoidal grating, 2.6° in diameter, 2.3 cycles per degree, contrast 90%, positioned at 3.2° eccentricity in the lower visual field. The phase was randomized so that orientation was the only cue that could be used to solve the task. A trial was initiated by the onset of a fixation point, which the subject had to fixate within a $1 \times 1^\circ$ window for 300 ms. The subject had to hold fixation for another 250 ms while a grating was shown. Only one orientation, either tilted clockwise or anticlockwise with respect to the oblique reference orientation, was shown on each trial. After the stimulus was switched off, two response targets appeared above and below the stimulus position on which the subject had to make a saccadic eye movement to the correct target to receive apple juice. Once the subject had learned the task, training proceeded using a transformed up–down staircase procedure²⁹, which converged on an orientation difference corresponding to an 84% correct criterion. Thresholds were defined as the logarithmic mean of the reversal points. The monkeys performed between 2,000 and 5,000 trials during each daily session.

Electrophysiology

We made extracellular recordings transdurally. The position of the receptive field was determined using multicellular activation by a small, flickering dot. When the response of a single cell was isolated, the optimal phase was determined and the response to 16 different orientations was recorded (15–30 trials per orientation). Spikes were counted in two 300-ms intervals, one before stimulus onset (baseline) and one starting 40 ms after stimulus onset (response).

We compared three populations of neurons: (1) trained neurons ($n = 650$); (2) passively stimulated neurons ($n = 273$); and (3) naive neurons ($n = 507$). Analysis of variance was applied to test for a significant response, and for an effect of orientation. From the 1,430 responsive single cells, 1,334 (93%) were orientation tuned. Cells were assigned to supragranular, infragranular or granular layers on the basis of the spontaneous activity

and depth of the electrode³⁰. Cells from layers 2–3 and 5–6 combined comprised nearly 80% of the total (487 trained, 383 naive and 199 passively stimulated neurons). Orientation tuning curves were normalized to their maximum and fitted using a polynomial of the 10th order. The squared sum of errors of the fitting averaged only 0.06. We omitted 7% (trained cells) to 10% (naive cells) of the cells because of a poor fit (squared sum of errors >0.2). We took the maximum of this fitted curve as the cell's preferred orientation and the tangent to the curve as the slope at the trained orientation, or at the untrained oblique orientation (NTO). Neurons were divided into 8-degree-wide groups according to the angle between their preferred orientation and trained orientation (or NTO). The numbers of neurons in each preferred orientation group are indicated between brackets in Fig. 3c. The slope changes derived from the fitted curves were independent of the fitting method, as similar results were obtained using a different order polynomial (8 or 12) or a spline fit.

Neurometric performance was determined using Bayes's rule^{22,23}. Performance of an ideal classifier was measured by computing the probability that one of two orientations was presented, given the responses of a set of neurons. This set was randomly selected either from all cells with preferred orientations within 45° of the trained orientation (naive $n = 175$; trained $n = 230$), from all cells with a preferred orientation within 45° of the NTO (naive $n = 170$; trained $n = 224$), or from each preferred orientation group independently. Mean proportions of correct response are averages of 30 computations, each made on the basis of a new selection of cells. Each computation was based on 500 trials, taken from a Poisson distribution derived from the mean firing rate of the cell. Mean responses to orientations, presented in 1° steps, were taken from the normalized fitted polynomials and multiplied by the average maximum rate of that group of cells. Standard error of the mean of the per cent that were correct varied between 0.1–1.2%. The values for per cent correct were z-transformed. After linear regression, the 84% threshold corresponds to the standard deviation (z-score of 1).

Received 21 March; accepted 5 June 2001.

1. Ahissar, M. & Hochstein, S. Task difficulty and the specificity of perceptual learning. *Nature* **387**, 401–406 (1997).
2. Crist, R. E., Kapadia, M. K., Westheimer, G. & Gilbert, C. D. Perceptual learning of spatial localization: specificity for orientation, position, and context. *J. Neurophysiol.* **78**, 2889–2894 (1997).
3. Fiorentini, A. & Berardi, N. Perceptual learning specific for orientation and spatial frequency. *Nature* **287**, 43–44 (1980).
4. Matthews, N., Liu, Z., Geesaman, B. J. & Qian, N. Perceptual learning on orientation and direction discrimination. *Vision Res.* **39**, 3692–3701 (1999).
5. Poggio, T., Fahle, M. & Edelman, S. Fast perceptual learning in visual hyperacuity. *Science* **256**, 1018–1021 (1992).
6. Vogels, R. & Orban, G. A. The effect of practice on the oblique effect in line orientation judgements. *Vision Res.* **25**, 1679–1687 (1985).
7. Schoups, A. A., Vogels, R. & Orban, G. A. Human perceptual learning in identifying the oblique orientation: retinotopy, orientation specificity and monocularly. *J. Physiol.* **483**, 797–810 (1995).
8. Schoups, A. A. & Orban, G. A. Interocular transfer in perceptual learning of a pop-out discrimination task. *Proc. Natl Acad. Sci. USA* **93**, 7358–7362 (1996).
9. Recanzone, G. H., Merzenich, M. M., Jenkins, W. M., Grajski, K. A. & Dinse, H. R. Topographic reorganization of the hand representation in cortical area 3b of owl monkeys trained in a frequency-discrimination task. *J. Neurophysiol.* **67**, 1031–1056 (1992).
10. Recanzone, G. H., Schreiner, C. E. & Merzenich, M. M. Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *J. Neurosci.* **13**, 87–103 (1993).
11. Nudo, R. J., Milliken, G. W., Jenkins, W. M. & Merzenich, M. M. Use-dependent alterations of movement representations in primary motor cortex of adult squirrel monkeys. *J. Neurosci.* **16**, 785–807 (1996).
12. Kaas, J. H. et al. Reorganization of retinotopic cortical maps in adult mammals after lesions of the retina. *Science* **248**, 229–231 (1990).
13. Gilbert, C. D. & Wiesel, T. N. Receptive field dynamics in adult primary visual cortex. *Nature* **356**, 150–152 (1992).
14. Zohary, E., Celebrini, S., Britten, K. H. & Newsome, W. T. Neuronal plasticity that underlies improvement in perceptual performance. *Science* **263**, 1289–1292 (1994).
15. Zohary, E. & Newsome, W. T. Perceptual learning in a direction discrimination task is not based upon enhanced neuronal sensitivity in the STS. *Invest. Ophthalmol. Vis. Sci.* **35**, 1663 (1994).
16. Regan, D. & Beverley, K. I. Postadaptation orientation discrimination. *J. Opt. Soc. Am. A* **2**, 147–155 (1985).
17. Bradley, A., Skottun, B. C., Ohzawa, I., Sclar, G. & Freeman, R. D. Visual orientation and spatial frequency discrimination: a comparison of single neurons and behavior. *J. Neurophysiol.* **57**, 755–772 (1987).
18. Vogels, R. & Orban, G. A. How well do response changes of striate neurons signal differences in orientation: a study in the discriminating monkey. *J. Neurosci.* **10**, 3543–3558 (1990).
19. Qian, N. & Matthews, N. A physiological theory for visual perceptual learning of orientation discrimination. *Soc. Neurosci. Abs.* **25**, 1316 (1999).
20. Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A. C. & Suarez, H. H. Recurrent excitation in neocortical circuits. *Science* **269**, 981–985 (1995).
21. Somers, D. C., Nelson, S. B. & Sur, M. An emergent model of orientation selectivity in cat visual cortical simple cells. *J. Neurosci.* **15**, 5448–5465 (1995).
22. Zhang, K., Ginzburg, I., McNaughton, B. & Sejnowski, T. J. Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *J. Neurophysiol.* **79**, 1017–1044 (1998).
23. Oram, M. W., Foldiak, P., Perrett, D. I. & Sengpiel, F. The 'ideal homunculus': decoding neural population signals. *Trends Neurosci.* **21**, 259–265 (1998).
24. Panzeri, S., Schultz, S. R., Treves, A. & Rolls, E. T. Correlations and the encoding of information in the nervous system. *Proc. R. Soc. Lond. B* **266**, 1001–1012 (1999).
25. Recanzone, G. H., Merzenich, M. M. & Schreiner, C. E. Changes in the distributed temporal response properties of SI cortical neurons reflect improvements in performance on a temporally based tactile discrimination task. *J. Neurophysiol.* **67**, 1071–1091 (1992).

26. Doshier, B. A. & Lu, Z. Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proc. Natl Acad. Sci. USA* **95**, 13988–13993 (1998).
27. Gold, J., Bennett, P. J. & Sekuler, A. B. Signal but not noise changes with perceptual learning. *Nature* **402**, 176–178 (1999).
28. Judge, S. J., Richmond, B. J. & Chu, F. C. Implantation of magnetic search coils for measurement of eye position: an improved method. *Vision Res.* **20**, 535–538 (1980).
29. Wetherill, G. B. & Levitt, H. Sequential estimation of points on a psychometric function. *Brit. J. Math. Stat. Psychol.* **18**, 1–10 (1965).
30. Snodderly, D. M. & Gur, M. Organization of striate cortex of alert, trained monkeys (*Macaca fascicularis*): ongoing activity, stimulus selectivity, and widths of receptive field activating regions. *J. Neurophysiol.* **74**, 2100–2125 (1995).

Acknowledgements

We would like to thank K. Claeys, P. Janssen, Z. Li, H. Op de Beeck, H. Peuskens, S. Raiguel, N. Sachs and W. Vanduffel for critical discussions, and M. DePaep, P. Kayenbergh, G. Meulemans, G. Vanparijs for technical assistance. A.S. is supported by a fellowship from FWO. This project was funded by grants from FWO (A.S.) GSKE (R.V.) NSF and NIH (N.Q.) and from DWTC (G.O.)

Correspondence and requests for materials should be addressed to A.S. (e-mail: annick.schoups@med.kuleuven.ac.be).

Regulation of DNA replication fork progression through damaged DNA by the Mec1/Rad53 checkpoint

José Antonio Tercero & John F. X. Diffley

Imperial Cancer Research Fund, Clare Hall Laboratories, South Mimms EN6 3LD, UK

The checkpoint kinase proteins Mec1 and Rad53 are required in the budding yeast, *Saccharomyces cerevisiae*, to maintain cell viability in the presence of drugs causing damage to DNA or arrest of DNA replication forks^{1–3}. It is thought that they act by inhibiting cell cycle progression, allowing time for DNA repair to take place. Mec1 and Rad53 also slow S phase progression in response to DNA alkylation⁴, although the mechanism for this and its relative importance in protecting cells from DNA damage have not been determined. Here we show that the DNA-alkylating agent methyl methanesulphonate (MMS) profoundly reduces the rate of DNA replication fork progression; however, this moderation does not require Rad53 or Mec1. The accelerated S phase in checkpoint mutants⁴, therefore, is primarily a consequence of inappropriate initiation events^{5–7}. Wild-type cells ultimately complete DNA replication in the presence of MMS. In contrast, replication forks in checkpoint mutants collapse irreversibly at high rates. Moreover, the cytotoxicity of MMS in checkpoint mutants occurs specifically when cells are allowed to enter S phase with DNA damage. Thus, preventing damage-induced DNA replication fork catastrophe seems to be a primary mechanism by which checkpoints preserve viability in the face of DNA alkylation.

MMS slows S phase progression in checkpoint-proficient yeast strains. This might be attributed entirely to the inhibition of late origin firing in these cells^{5,6}. Alternatively, a reduced rate of replication fork progression ('fork rate') may also contribute to the slow S phase. Likewise, the acceleration through S phase in MMS seen in checkpoint mutants may solely be a consequence of their inability to inhibit late origin firing or may be an indication that checkpoints also regulate fork rates. To follow replication forks, we used a density transfer approach (refs 8 and 9 and Supplementary Information). Cells were grown in 'heavy' isotopes (¹³C glucose, ¹⁵N ammonium sulphate) to ensure full isotope substitution of the parental DNA. These cells were arrested in G1 phase with α -factor

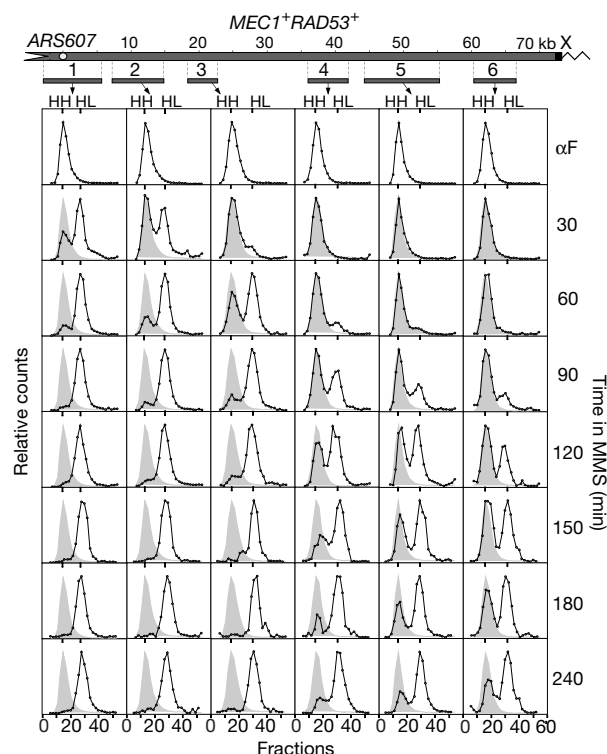


Figure 1 Replication fork progression in a checkpoint-proficient strain. A time course of DNA replication in part of the right arm of chromosome VI in YJT85 was analysed after release from α -factor (α F) arrest into medium with 0.033% MMS by dense-isotope transfer, as described in the text. The positions of potential replication origins and of the six *Clal*–*Sall* restriction fragments examined are shown at the top. The position of unreplicated heavy–heavy (HH) and fully replicated heavy–light (HL) peaks is indicated. At later points the position of the initial HH peak is shown for comparison (grey area). The replication pattern in this *smf1* Δ mutant is virtually identical to that seen in an *SML1*⁺ strain (see Supplementary Information).

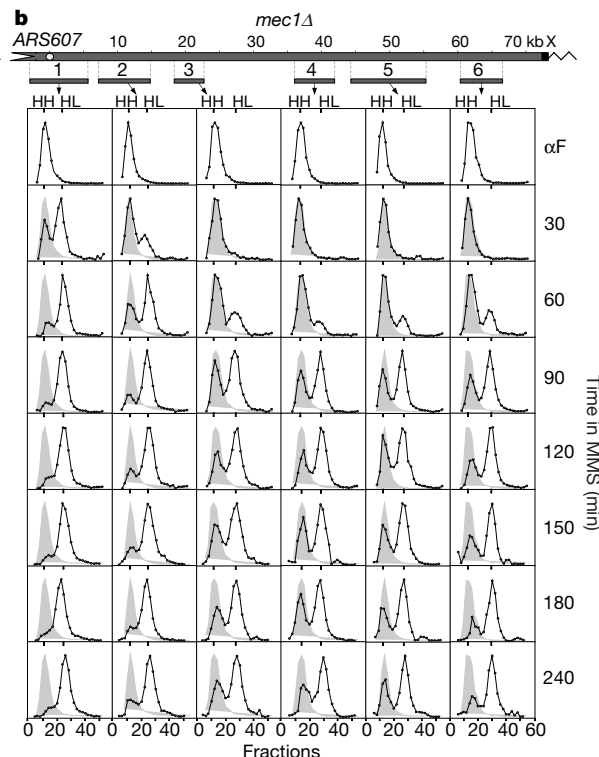
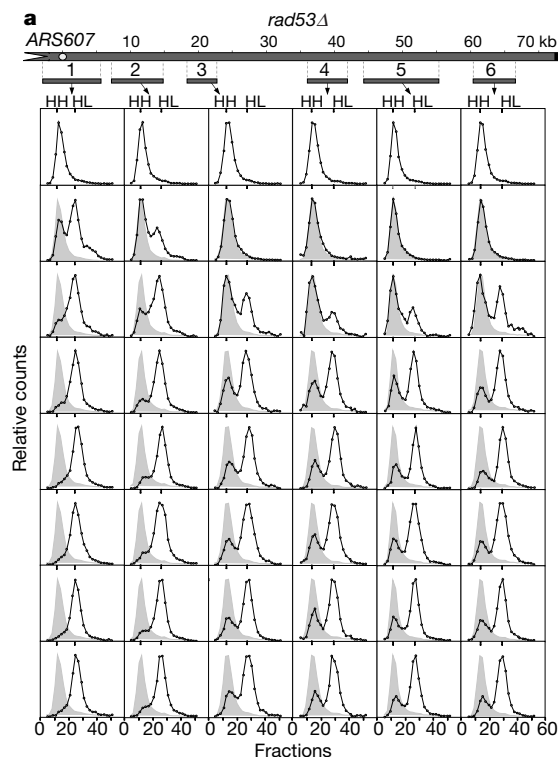


Figure 2 Replication fork progression in checkpoint-deficient strains. Replication in YJT81 (a) and YJT82 (b) strains with MMS was followed exactly as in Fig. 1. Strains harbour deletions of *RAD53* (a) or *MEC1* (b). Both strains are kept alive by deletion of

mating pheromone and transferred to medium containing 'light' isotopes (¹²C glucose, ¹⁴N ammonium sulphate). Cells were then released from the α -factor arrest into fresh medium containing light isotopes either with or without MMS. At indicated time points, DNA was digested with restriction enzymes and separated on caesium chloride gradients. The position of individual restriction fragments in the gradient was determined by DNA slot blot hybridization. DNA replication of a restriction fragment is seen by its transfer from the heavy–heavy to the heavy–light peak.

We followed the replication of six restriction fragments in a replicon at the end of chromosome VI (refs 10 and 11). At the top of Fig. 1, the main features of this replicon are illustrated. *ARS607* (autonomously replicating sequence 607) is an efficient, early-firing replication origin. The right arm of chromosome VI, like all yeast chromosomes, contains a sub-telomeric X element. X elements have ARS activity on plasmids¹², although it is not known whether the X element on chromosome VI is an active origin. However, to avoid complications from telomere dysfunction in checkpoint mutants^{13,14}, we did not remove this element. Two inefficient, later-firing origins, *ARS608* and *ARS609*, were deleted to allow us to follow replication forks between *ARS607* and the X-element ARS across a region of 75 kilobases (kb), about twice the average inter-origin distance in yeast.

Figure 1 shows DNA replication in the checkpoint-proficient strain in the presence of MMS. In the α -factor-arrested cells (top row), all of the restriction fragments are present in the heavy–heavy peak. Rows 2–8 illustrate several points about DNA replication after release from α -factor arrest in the presence of MMS when the *Mec1*/*Rad53* checkpoint is intact. First, replication is very slow. Even by the end of the experiment (240 min), fragments 4–6 have not been completely replicated in all cells. The average fork rate across this replicon was calculated to be about 300 bp min^{−1} (see Supplementary Information), 5–10 times lower than fork rates in the absence of MMS^{8,15}. Second, because density substitution is a quantitative assay and because replication proceeds with reasonable synchrony, we can conclude that the entire region is replicated from left to right:

SML1, which suppresses the lethality of *RAD53* and *MEC1* deletion without suppressing their checkpoint defect²². HH, heavy–heavy peak; HL, heavy–light peak.

fragment 1 replicates before fragment 2, fragment 2 before 3, and so on. This is consistent with efficient activation of *ARS607* in *MMS*⁵ and also indicates that the origin associated with the X element makes little or no contribution to the replication of this region, suggesting that it is not activated in this experiment. Third, replication forks remain active throughout the entire experiment. For example, there is significant replication of fragments 4–6 even between the last two time points (180 and 240 min). Finally, replication even across this long replicon will, ultimately, be completed. At the end of the experiment, virtually all of fragments 1–3 are in the heavy–light peak, while the continuing replication of fragments 4–6 between the last time points suggests that they, too, should ultimately be completely replicated. Thus, the slow S phase progression seen in *MMS* in checkpoint-proficient cells is due to a combination of slow but stable replication forks and checkpoint-dependent origin inhibition^{5,6}.

The replication of this region of chromosome VI in the *rad53* and *mec1* checkpoint mutants was determined in parallel. This experiment (Fig. 2) illustrates several similarities to, and differences from, the pattern shown in Fig. 1. First, as in the wild-type strain, replication from left to right can clearly be seen in the left half of this replicon in both mutants: fragment 1 replicates before fragment 2, and fragment 2 replicates before fragment 3. Second, the progression of this replication fork in both mutants is slow. The average fork rate across the first 20 kb of this replicon was approximately 300 bp min⁻¹ (see Supplementary Information) in both mutants, the same as the wild-type strain. In contrast to the wild type, however, replication forks do not proceed from left to right across the entire region. Instead, forks seem to initiate some time around 60 min from the right end of the chromosome and proceed leftward. This can be seen in the *rad53* mutant at the 60-min time point, where fragment 6 has replicated before fragments 4 and 5. Fork direction is more difficult to deduce in fragments 4 and 5, which seem to replicate at similar times in the checkpoint mutants. This is presumably because, in some cells, these fragments are replicated from a rightward fork and in other cells from a leftward fork. It is unlikely that this is due to origin activation within fragments 4 or 5, as this region contains no known ARS elements¹⁶. These results indicate that the X-element-associated ARS becomes activated in these cells, consistent with the fact that these ARSs are competent to be activated in their chromosomal location¹⁷ and with the previously established role for the checkpoint kinases in preventing the activation of late-firing^{5,6} and dormant⁷ replication origins.

Finally, one of the most striking features of replication in the two checkpoint mutants is that a significant fraction of replication forks arrest before replication of the entire replicon is completed. In sharp contrast to the checkpoint-proficient strain, in which replication continues to the end of the experiment, there is no further DNA synthesis after 120 min in either of the checkpoint mutants. However, in both mutants, significant amounts of DNA remain unreplicated. Figure 2 shows that the amount of unreplicated DNA increases with increasing distance from the origins. In both mutants, there is more unreplicated DNA in fragment 3 than in fragment 2 and more in fragment 2 than in fragment 1. Therefore, in both checkpoint mutants, replication forks terminate irreversibly and apparently randomly at about 2% per kb across the entire region (see Supplementary Information for more detail).

Replication forks in yeast must travel, on average, 20 kb. About 40% of the replication forks terminate before 20 kb in both mutants (Fig. 2 and Supplementary Information). The probability, then, that two converging forks will both terminate prematurely, and thereby prevent complete replication of the replicon is roughly $0.4 \times 0.4 = 0.16$, or 16%. Thus, assuming there are about 400 replicons in yeast¹⁸, there will be about 64 replicons in which both forks have stopped before normal fork convergence. Given that a single such event should be lethal, fork catastrophe could easily account for the loss of viability in these mutants. Because these

replicons will have replicated at least some of their DNA and because the remaining 84% of the replicons will complete replication, S phase appears to be complete by flow cytometry (ref. 4 and Supplementary Information). The more sensitive assay used here, however, shows that DNA replication is never completed in the checkpoint mutants.

If fork breakdown is responsible for the cytotoxicity of *MMS* in checkpoint mutants, the lethal effects of *MMS* should be confined to S phase. To test this, we examined the viability of checkpoint mutant cells treated with *MMS* either while held in G1 with α -factor or after release from α -factor. In both cases, cells were exposed to different amounts of *MMS* for 1 h. Figure 3a shows that when *mec1* or *rad53* mutant cells were exposed to *MMS* during G1 arrest, they were significantly more resistant to cell death than cells allowed to

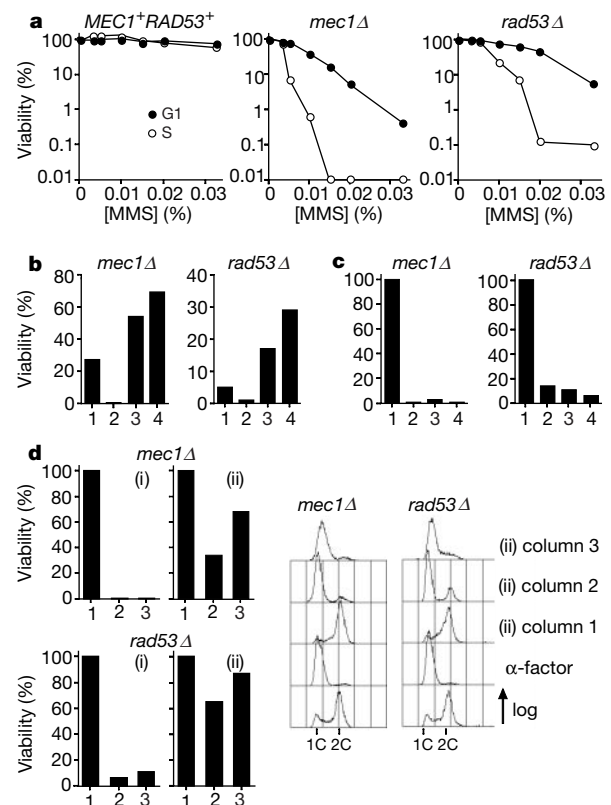


Figure 3 Cytotoxicity of *MMS* in checkpoint mutants occurs during S phase.

a, Checkpoint-proficient and -deficient strains (YJT85, YJT81 and YJT82) were treated with the indicated concentrations of *MMS* for 1 h either while held in α -factor arrest (G1) or after release from α -factor arrest (S phase) and viability was determined. **b**, Checkpoint mutant cells were treated with *MMS* for 1 h and tested for viability. Columns: 1, cells were held in α -factor during treatment with *MMS* and plated immediately for viability; 2, cells were treated with *MMS* after release from α -factor and plated immediately; 3 and 4, cells were held in α -factor while treated with *MMS* then transferred to fresh medium containing α -factor but lacking *MMS* for 1 h (column 3) or 2 h (column 4) before plating for viability. *MMS* concentration was 0.010% for *smi1Δmec1Δ* and 0.033% for *smi1Δrad53Δ*. **c**, Treatments for checkpoint mutant cells. Columns: 1, cells were plated after α -factor arrest; 2, cells were released from α -factor arrest, treated with 0.010% *MMS* for 1 h and plated; 3 and 4, cells were released from α -factor, treated with *MMS*, then transferred to fresh medium lacking *MMS* but containing 5 μ g ml⁻¹ nocodazole for 1 h (column 3) or 2 h (column 4). **d**, A culture of α -factor-arrested checkpoint mutant cells was divided in half and released from α -factor arrest. One half (i) was treated immediately with 0.010% *MMS* and transferred after 75 min to fresh medium lacking *MMS* but containing α -factor. The other half (ii) was treated 75 min after release with 0.010% *MMS* for 60 min then transferred to fresh medium lacking *MMS* but containing α -factor. Cells were plated for viability before *MMS* addition (column 1), immediately after *MMS* treatment (column 2) and 2 h (i) or 1 h (ii) after transfer to fresh medium (column 3). DNA content is shown for the indicated time points.

progress through S phase in the presence of MMS. We repeatedly saw that *mec1* mutants are more sensitive to MMS than *rad53* mutants in both G1 and S phase. Similarly, *mec1* mutants show more fork catastrophe than *rad53* mutants (Fig. 2). As both are deletion mutants, partial gene activity cannot account for this difference, suggesting that Mec1 has a Rad53-independent as well as a Rad53-dependent role in preventing fork catastrophe. Both *mec1* and *rad53* mutants lose more viability than wild-type cells, even when held in G1 (Fig. 3a). To test whether this is due to residual DNA damage that kills cells when they pass through S phase after plating, we examined whether holding cells for additional time in the α -factor block after MMS treatment could rescue viability. Figure 3b shows that holding both *rad53* and *mec1* mutants in α -factor for an additional 1 or 2 h after MMS treatment resulted in significant increase in survival. From this we infer that DNA damage incurred outside of S phase, which can, apparently, be repaired in the checkpoint mutants, is lethal when these mutants enter S phase. Although this is consistent with the idea that checkpoints act by preventing cell cycle progression and, thus, allowing time for DNA repair, we note that wild-type cells released from α -factor into MMS do not show significant delays in either bud emergence (data not shown) or entry into S phase (Fig. 1 and Supplementary Information). To ensure that cell killing after release from α -factor occurred during S phase rather than during mitosis, we performed two experiments. First, after releasing cells from α -factor arrest for 1 h in the presence of MMS, we transferred them to nocodazole-containing medium for 1 or 2 h to delay entry into mitosis. This did not promote any significant increase in survival (Fig. 3c). Finally, we released cells from α -factor arrest for 75 min before adding MMS for 1 h. At 75 min, most cells have 2C DNA content and are uninucleate (Fig. 3d and data not shown), indicating that they have completed S phase but have not yet entered mitosis. Figure 3d shows that, if MMS is added after S phase is complete, the checkpoint mutants have much higher viability than they do when MMS is added before S phase. Taken together, these experiments demonstrate that it is passage through S phase in the presence of DNA damage that kills the checkpoint mutant cells.

Our experiments define the parameters of the intra-S checkpoint with respect to DNA replication. First, MMS treatment dramatically reduces the rate at which replication forks proceed. Second, this reduction in replication fork rate is independent of Mec1 and Rad53. It is, therefore, unlikely to be a checkpoint phenomenon and more likely to represent a physical impediment of replication fork progression by either DNA alkylation or some intermediate in lesion processing. Third, additional replication origins, which are not active in wild-type cells, become activated in *mec1* and *rad53* checkpoint mutants. These origins probably include the sub-telomeric X ARS as well as the normally inactive origins on chromosome III (ref. 7 and data not shown). Because the checkpoint mutants do not show accelerated rates of replication fork progression, the faster S phase seen in checkpoint mutants in MMS must be primarily due to inappropriate origin firing.

In addition to defining the parameters of the intra-S checkpoint, our experiments make several points about DNA replication through damaged DNA. Although replication forks move slowly in wild-type cells, they continue to progress, ultimately resulting in the complete replication of even a fairly long replicon such as the one examined on chromosome VI. One model for lesion-bypass synthesis involves a mechanism in which strand switching allows one nascent strand to act as a template for the other nascent strand¹⁹. We note that such an event would generate light–light DNA, which was never seen in either checkpoint-proficient or -deficient cells, suggesting that, if such a mechanism is used, it must involve only short tracts of strand switching. Most significantly, our experiments show that replication forks terminate irreversibly at a high rate in *rad53* and *mec1* checkpoint mutants. Moreover, the extreme cytotoxicity of MMS in checkpoint mutants requires passage through S

phase. Damage-induced DNA replication fork catastrophe, therefore, appears to be a major reason for the very high lethality of MMS in checkpoint mutants.

Aberrant DNA structures induced by MMS in transformed Chinese hamster ovary (CHO) cells are generated specifically during S phase²⁰, and early S phase replication patterns in cells blocked with aphidicolin are unstable when treated with checkpoint kinase inhibitors²¹. This suggests that checkpoint kinases may also act to aid replication through damaged DNA and replication fork blocks in mammalian cells. Many anti-cancer drugs act by damaging DNA or otherwise interfering with DNA replication. Our results may have implications for how these therapies should be delivered to checkpoint-defective tumour cells. □

Methods

Strains used were YJT80 (*ARS608Δ::HIS3*, *ARS609Δ::TRP1*, *ARS305Δ::kanMX*, *ade2-1::ADE2*, W303-1a background), YJT81 (*sm1Δ::URA3*, *rad53Δ::LEU2*, *ARS608Δ::HIS3*, *ARS609Δ::TRP1*, *ARS305Δ::kanMX*, *ade2-1::ADE2*, W303-1a background), YJT82 (*sm1Δ::URA3*, *mec1Δ::LEU2*, *ARS608Δ::HIS3*, *ARS609Δ::TRP1*, *ARS305Δ::kanMX*, *ade2-1::ADE2*, W303-1a background) and YJT85 (*sm1Δ::URA3*, *ARS608Δ::HIS3*, *ARS609Δ::TRP1*, *ARS305Δ::kanMX*, *ade2-1::ADE2*, W303-1a background). A list of oligonucleotides used to construct these strains can be found in the Supplementary Information.

Flow cytometry and density transfer were performed and analysed essentially as described (ref. 9 and http://fangman-brewer.genetics.washington.edu/density_transfer.html). DNA was digested with *Clal* and *Sall* before gradient centrifugation in caesium chloride. DNA probes for slot blot hybridization were amplified by polymerase chain reaction (PCR). Probe number corresponds to fragment number and were as follows: probe 1, nucleotides 198945–199832; probe 2, nucleotides 211014–211996; probe 3, nucleotides 218011–218700; probe 4, nucleotides 240009–240679; probe 5, nucleotides 243315–244200; and probe 6, nucleotides 260048–261088.

Amounts of DNA replication and fork rates were determined essentially as described by Reynolds *et al.*⁸, in Supplementary Information and at http://fangman-brewer.genetics.washington.edu/density_transfer.html.

Viability was determined after dilution and sonication by plating about 400 cells in duplicate onto YPD plates.

Received 10 May; accepted 21 June 2001.

- Lowndes, N. F. & Murguia, J. R. Sensing and responding to DNA damage. *Curr. Opin. Genet. Dev.* **10**, 17–25 (2000).
- Zhou, B. B. & Elledge, S. J. The DNA damage response: putting checkpoints in perspective. *Nature* **408**, 433–439 (2000).
- Rhind, N. & Russell, P. Checkpoints: it takes more than time to heal some wounds. *Curr. Biol.* **10**, R908–R911 (2000).
- Paulovich, A. G. & Hartwell, L. H. A checkpoint regulates the rate of progression through S phase in *S. cerevisiae* in response to DNA damage. *Cell* **82**, 841–847 (1995).
- Shirahige, K. *et al.* Regulation of DNA-replication origins during cell-cycle progression. *Nature* **395**, 618–621 (1998).
- Santocanale, C. & Diffley, J. F. X. A Mec1- and Rad53-dependent checkpoint controls late-firing origins of DNA replication. *Nature* **395**, 615–618 (1998).
- Santocanale, C., Sharma, K. & Diffley, J. F. X. Activation of dormant origins of DNA replication in budding yeast. *Genes Dev.* **13**, 2360–2364 (1999).
- Reynolds, A. E., McCarroll, R. M., Newlon, C. S. & Fangman, W. L. Time of replication of ARS elements along yeast chromosome III. *Mol. Cell. Biol.* **9**, 4488–4494 (1989).
- Tercero, J. A., Labib, K. & Diffley, J. F. X. DNA synthesis at individual replication forks requires the essential initiation factor, Cdc45p. *EMBO J.* **19**, 2082–2093 (2000).
- Yamashita, M. *et al.* The efficiency and timing of initiation of replication of multiple replicons of *Saccharomyces cerevisiae* chromosome VI. *Genes Cells* **2**, 655–665 (1997).
- Friedman, K. L., Brewer, B. J. & Fangman, W. L. Replication profile of *Saccharomyces cerevisiae* chromosome VI. *Genes Cells* **2**, 667–678 (1997).
- Chan, C. S. & Tye, B. K. Organization of DNA sequences and replication origins at yeast telomeres. *Cell* **33**, 563–573 (1983).
- Craven, R. J. & Petes, T. D. Involvement of the checkpoint protein Mec1p in silencing of gene expression at telomeres in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **20**, 2378–2384 (2000).
- Longhese, M. P., Paciotti, V., Neeke, H. & Lucchini, G. Checkpoint proteins influence telomeric silencing and length maintenance in budding yeast. *Genetics* **155**, 1577–1591 (2000).
- Rivin, C. J. & Fangman, W. L. Replication fork rate and origin activation during the S phase of *Saccharomyces cerevisiae*. *J. Cell Biol.* **85**, 108–115 (1980).
- Shirahige, K., Iwasaki, T., Rashid, M. B., Ogasawara, N. & Yoshikawa, H. Location and characterization of autonomously replicating sequences from chromosome VI of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **13**, 5043–5056 (1993).
- Stevenson, J. B. & Gottschling, D. E. Telomeric chromatin modulates replication timing near chromosome ends. *Genes Dev.* **13**, 146–151 (1999).
- Campbell, J. L. & Newlon, C. S. in *The Molecular Biology of the Yeast Saccharomyces; Genome Dynamics, Protein Synthesis and Energetics* (eds Broach, J. R., Pringle, J. R. & Jones, E. W.) 41–146 (Cold Spring Harbor Press, Cold Spring Harbor, 1991).
- Higgins, N. P., Kato, K. & Strauss, B. A model for replication repair in mammalian cells. *J. Mol. Biol.* **101**, 417–425 (1976).
- Schwartz, J. L. Monofunctional alkylating agent-induced S-phase-dependent DNA damage. *Mutat. Res.* **216**, 111–118 (1989).

21. Dimitrova, D. S. & Gilbert, D. M. Temporally coordinated assembly and disassembly of replication factories in the absence of DNA synthesis. *Nature Cell Biol.* **2**, 686–694 (2000).
22. Zhao, X., Muller, E. G. & Rothstein, R. A suppressor of two essential checkpoint genes identifies a novel protein that negatively affects dNTP pools. *Mol. Cell* **2**, 329–340 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank past and present members of the laboratory for discussions and K. Labib and D. G. Quintana for critical reading of the manuscript. We also thank M. Foiani and colleagues for discussions and for communicating unpublished results. This work was supported by the Imperial Cancer Research Fund. J.A.T. was supported by a Human Frontier Science Program fellowship.

Correspondence and requests for materials should be addressed to J.F.X.D. (e-mail: j.diffley@icrf.icnet.uk).

The DNA replication checkpoint response stabilizes stalled replication forks

Massimo Lopes*, Cecilia Cotta-Ramusino*, Achille Pelliccioli*, Giordano Liberi*, Paolo Plevani*, Marco Muzi-Falconi*, Carol S. Newlon† & Marco Foiani*

* Istituto F.I.R.C. di Oncologia Molecolare, Via Serio 21, 20141, Milano, Italy, and Dipartimento di Genetica e di Biologia dei Microrganismi, Università degli Studi di Milano, Via Celoria 26, 20133, Milano, Italy

† Department of Microbiology & Molecular Genetics, UMDNJ—New Jersey Medical School, Newark, New Jersey 07103, USA

In response to DNA damage and blocks to replication, eukaryotes activate the checkpoint pathways that prevent genomic instability and cancer by coordinating cell cycle progression with DNA repair^{1–5}. In budding yeast, the checkpoint response requires the

Mec1-dependent activation of the Rad53 protein kinase^{3,4,6}. Active Rad53 slows DNA synthesis when DNA is damaged⁷ and prevents firing of late origins of replication^{8,9}. Further, *rad53* mutants are unable to recover from a replication block¹⁰. Mec1 and Rad53 also modulate the phosphorylation state of different DNA replication and repair enzymes^{6,11–13}. Little is known of the mechanisms by which checkpoint pathways interact with the replication apparatus when DNA is damaged or replication blocked. We used the two-dimensional gel technique¹⁴ to examine replication intermediates in response to hydroxyurea-induced replication blocks. Here we show that hydroxyurea-treated *rad53* mutants accumulate unusual DNA structures at replication forks. The persistence of these abnormal molecules during recovery from the hydroxyurea block correlates with the inability to dephosphorylate Rad53. Further, Rad53 is required to properly maintain stable replication forks during the block. We propose that Rad53 prevents collapse of the fork when replication pauses.

Hydroxyurea, an inhibitor of ribonucleotide reductase, pauses replication by limiting dNTP pools¹⁵. Replication forks stall, leading to Rad53 phosphorylation and activation⁶. We analysed the replication intermediates at an early origin of replication (*ARS305*)¹⁶ in wild-type and *rad53* cells released from a G1 block in the presence of hydroxyurea (Fig. 1 and see Supplementary Information). In wild-type cells, bubbles and large Y-shaped molecules accumulated within 30–60 min of G1 release and started to decrease after 90 min. The bubble arc represents origins that have been fired bidirectionally, while large Y molecules result from asymmetric progression of replication forks out of the restriction fragment containing *ARS305* (305-rf). *rad53* cells accumulated bubbles with kinetics different from wild type: in general, the relative level of bubbles was lower and they started to decrease 30 min earlier than in wild-type cells (Fig. 1). Moreover, *rad53* cells accumulated small Y molecules and a cone-shaped signal resulting from molecules migrating similarly to double-Y- and/or X-shaped structures¹⁷ (arrows, Fig. 1), which persisted for at least 3 h. Hence, hydroxyurea-treated *rad53* mutants accumulated DNA structures (small Y molecules and a cone-shaped signal) throughout the treatment, concomitantly with a relative reduction of bubbles.

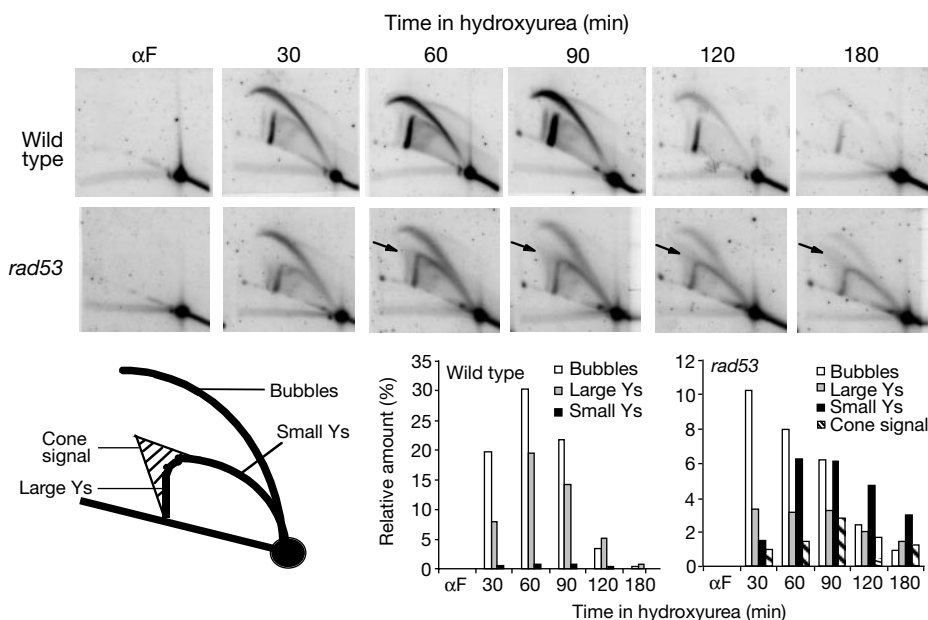


Figure 1 *rad53-K227A* mutant cells accumulate abnormal DNA structures at *ARS305* in response to hydroxyurea treatment. Wild-type W303-1A and CY2034 *rad53* mutant cells were grown in YPD medium, pre-synchronized by α -factor (α F) treatment and released from the G1 block in fresh medium containing 0.2 M hydroxyurea. DNA was prepared from cells collected at the indicated times: 10- μ g aliquots were cut with 100 units of

NcoI, electrophoresed as previously reported¹⁴, transferred to nylon membranes that were probed with a ³²P-labelled *Bam*HI-*NcoI* 3.0-kb fragment, spanning the *ARS305* origin. Arrows indicate the cone-shaped signal. A schematic representation of the replication intermediates discussed in the text and the relative quantification analysis are also presented.

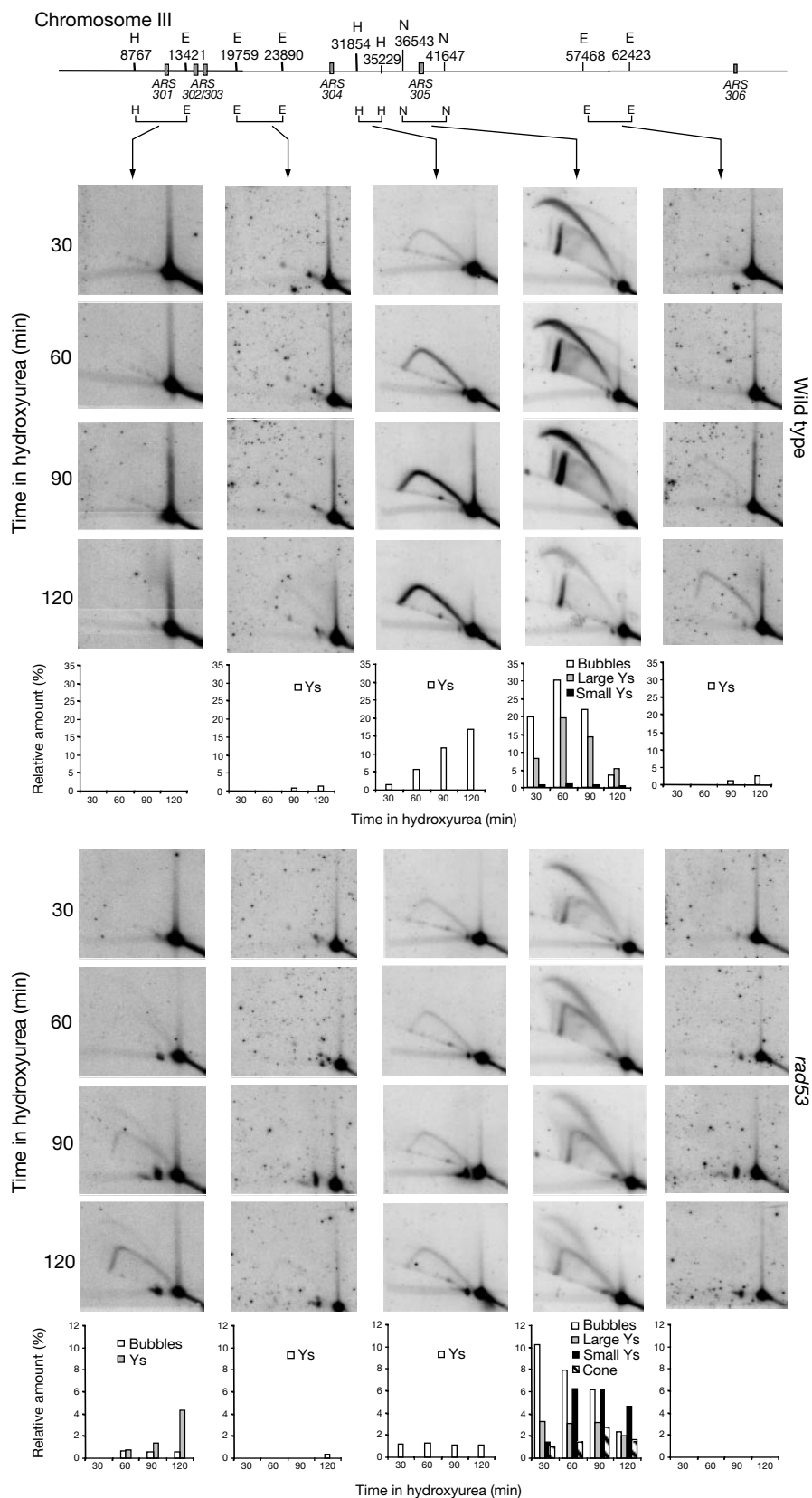


Figure 2 Analysis of replication intermediates in chromosome III regions adjacent to the *ARS305* origin. The same DNA preparation analysed in Fig. 1 was used to monitor replication intermediates in different regions of chromosome III, in wild-type and *rad53-K227A* cells, by cutting with the appropriate restriction enzymes to generate the

fragments drawn below the restriction map. H, *Hind*III; E, *Eco*RI; N, *Nco*I. Specific probes were obtained by polymerase-chain-reaction amplification with oligonucleotides complementary to the indicated regions. Quantification of replication intermediates is also presented.

We then monitored replication fork progression in hydroxyurea-treated wild-type and *rad53* cells by analysing different regions of chromosome III. Figure 2 shows that, in wild-type cells, replication intermediates initially accumulate in the 305-rf and then progressively invade chromosomal regions positioned 5 kilobases (kb) and 16 kb to the left, and 17 kb to the right, of *ARS305*. The replication intermediates that were first detected at 30 min in the region 5 kb to the left of the 305-rf represent the first leftward forks coming from *ARS305*. Accordingly, this signal gradually increased, indicating that the disappearance of replication intermediates in the 305-rf observed later is due to slow completion of replication even in the presence of hydroxyurea. In contrast, in hydroxyurea-treated *rad53* cells, we failed to detect significant accumulation of replication forks in regions flanking the 305-rf (Fig. 2), suggesting that fork progression was defective. In agreement with previous studies⁸, we found that *rad53* cells were unable to prevent firing of the dormant origin, *ARS301*, although its firing occurred well after *ARS305* activation (Fig. 2). The following observations rule out the possibility that in *rad53* cells the Y arc results from passive replication of the 305-rf: (1) bubbles appear before the Y arc in the 305-rf, indicating that the origin has been efficiently fired; (2) we did not observe accumulation of replication intermediates in *ARS305*-adjacent regions that could account for passive replication of the 305-rf; and (3) the *ARS305* Y arc is detectable earlier than replication intermediates in the *ARS301* restriction fragment.

The resetting of cell cycle progression during recovery from hydroxyurea is associated with Rad53 inactivation⁶ (Fig. 3a, b). We compared the resolution of replication intermediates during recovery in wild-type and *rad53* cells. Hydroxyurea-treated wild-type cells released from the hydroxyurea block under normal conditions were able to complete replication in 60 min (Fig. 3a). Thirty minutes after release from the hydroxyurea block, replication intermediates at *ARS305* were significantly reduced because

replication completed rapidly (Fig. 3c). At 90 min, bubbles and large Y molecules reappeared owing to a new round of replication (Fig. 3a, c). Hence in wild-type cells, bubbles and large Y molecules were rapidly resolved during recovery from hydroxyurea, indicating that stalled replication forks are very stable and can be reused after the block is released. This finding rules out the possibility that in wild-type cells stalled replication forks collapse and *de novo* firing of replicons is required to resume replication. Conversely, we found that *rad53* cells were unable to complete DNA replication after removal of hydroxyurea, and the mutant Rad53 protein remained phosphorylated (Fig. 3a, b). Two-dimensional gel analysis showed that, during the hydroxyurea treatment, *rad53* cells accumulated an intense signal of Y molecules that persisted throughout recovery (Fig. 3c). Hence, in response to hydroxyurea treatment, *rad53* mutants accumulate Y-shaped DNA structures that persisted during recovery, correlating with the inability to properly resume DNA replication. The maintenance of the hyperphosphorylated state of Rad53 mutant protein during recovery requires a functional Mec1 kinase: we found that in *mec1-td rad53* double mutants, released from hydroxyurea at 37 °C (*mec1-td* restrictive temperature), mutant Rad53 protein was rapidly dephosphorylated (data not shown). Hence, in *rad53* mutants, the Mec1-dependent checkpoint pathway leading to Rad53 phosphorylation is still sensing abnormal DNA structures even when hydroxyurea is removed from the medium. Intriguingly, it has been shown that Cds1, the *Schizosaccharomyces pombe* homologue of Rad53, is required to prevent the formation of DNA-damage structures during hydroxyurea-induced S-phase arrest¹⁸.

To test whether Rad53 was required to maintain stalled replication forks in a correct conformation in response to dNTP depletion, we first activated the checkpoint and then overexpressed a dominant negative version of Rad53, always in the presence of hydroxyurea (Fig. 4a and see Supplementary Information). The dominant

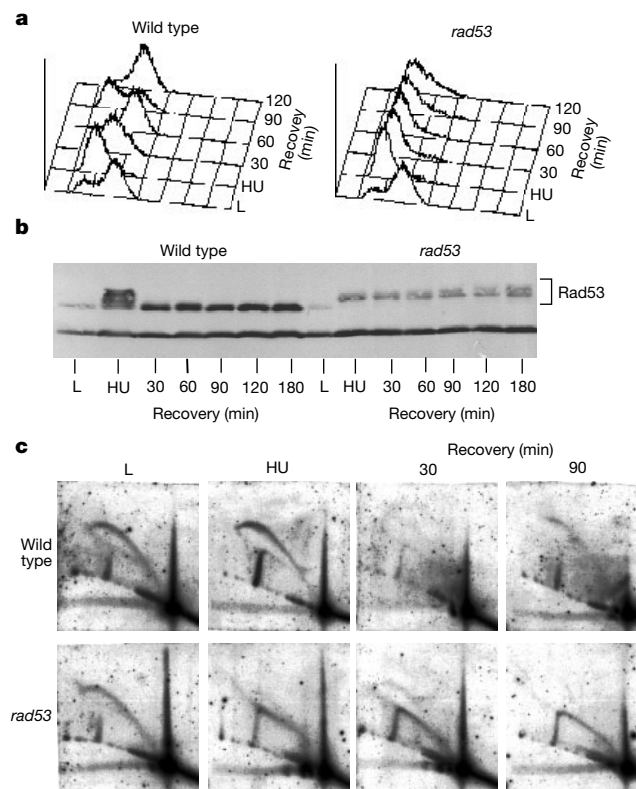


Figure 3 Y-shaped molecules in *rad53-K227A* mutant cells persist after hydroxyurea removal. **a**, Wild-type (W303-1A) and *rad53* mutant cells (CY2034), logarithmically growing in YPD medium at 28 °C (L), were first arrested by hydroxyurea treatment (HU, 0.2 M for 3 h) and then released from the hydroxyurea block in fresh medium lacking the

drug (recovery). The DNA content was analysed at the indicated times by FACS. **b**, Protein extracts were prepared at the indicated times, and analysed by western blotting using antibodies against Rad53. **c**, DNA samples were prepared at the indicated times and 10-μg aliquots were analysed by two-dimensional electrophoresis as described in Fig. 1.

negative Rad53 mutant protein can still be phosphorylated in *trans* but is unable to transduce the signal⁶ owing to a mutation that abolishes its kinase activity. *GAL1-rad53* cells were pre-synchronized in G1 and released from the G1 block in the presence of hydroxyurea under conditions that prevented *GAL1* expression. When more than 90% of the cells had large buds with early origins fired and replication forks stalled, we inactivated the checkpoint by overexpressing the dominant negative *rad53* allele. Figure 4a shows that, in control cells (vector), replication intermediates persisted throughout the hydroxyurea treatment (5.5 h total), gradually disappearing later (data not shown). In contrast to the control, as Rad53 function was impaired, the small Y molecules and the cone signal increased (Fig. 4a, arrows) at the expense of the accumulated bubbles and large Y molecules. Further, linear molecules accumulated, as deduced by the observation that the relative amount of total replication intermediates decreased with time (20.5% in hydroxyurea versus 10.7% at 120 min). We therefore conclude that *rad53* mutants are unable to properly maintain blocked replication forks in response to hydroxyurea treatment.

The following conclusions can be drawn from our results. First, hydroxyurea-treated wild-type cells fire early replication origins and use stalled forks to resume replication after the block is released. Second, *rad53* mutations do not affect the early steps of DNA synthesis in response to dNTP depletion. Third, *rad53* mutants are unable to maintain stalled replication forks in a proper conformation: *rad53* mutants first accumulate normal replication intermediates before other DNA structures, which probably prevent resumption of DNA synthesis when hydroxyurea is removed and contribute to maintaining phosphorylation of Rad53 protein. These observations are consistent with genetic data described for *S. pombe*^{18,19}.

The accumulation in hydroxyurea-treated *rad53* cells of abnormal DNA structures is probably the result of replication fork

collapse. We suggest that the checkpoint response maintains the integrity of stalled forks by preventing replisome disassembly (Fig. 4b). To test this hypothesis, we used temperature-sensitive replication mutants (*cdc17*, *pri1*, *pri2*) that cause replisome destabilization at restrictive temperature due to termolability of the encoded proteins. We found that mutant cells, released from G2 at 37 °C, arrested in S phase with active Rad53 and accumulated in the 305-rf small Y molecules and the cone signal, concomitantly with a reduction in the amount of bubbles (data not shown). Hence these mutants, despite exhibiting an active checkpoint, accumulated the same abnormal DNA structures as we observed in hydroxyurea-treated *rad53* mutants. Considering that the replisome itself is a target of the checkpoint response^{6,11}, and that checkpoint inhibitors affect the integrity of replication factories in S-phase arrested mammalian cells²⁰, perhaps Rad53 actively protects the stability of replication forks by modulating the phosphorylation state of crucial DNA replication proteins. Active Rad53 might also be required to restrain chromatin displacement ahead of stalled replication forks, preventing the accumulation of fragile single-stranded DNA. The small Y molecules, the complex structures that contribute to the cone signal and the linear molecules could arise from processing of unprotected bubbles, probably through unscheduled nucleolytic and/or recombination events, as has been shown in *Escherichia coli*^{21,22} (Fig. 4b). Indeed, small Y molecules can originate from broken bubbles^{23,24}. Different pathways could be responsible for fork processing in the absence of an active checkpoint as Rad53 activation leads to phosphorylation of several repair and recombination proteins. We found, however, that the Rad52-dependent recombination pathway does not contribute to the formation of abnormal structures in *rad53* mutant cells (data not shown).

In any case, the collapse of replication forks can easily account for the cell lethality observed in hydroxyurea-treated *rad53* cells and for

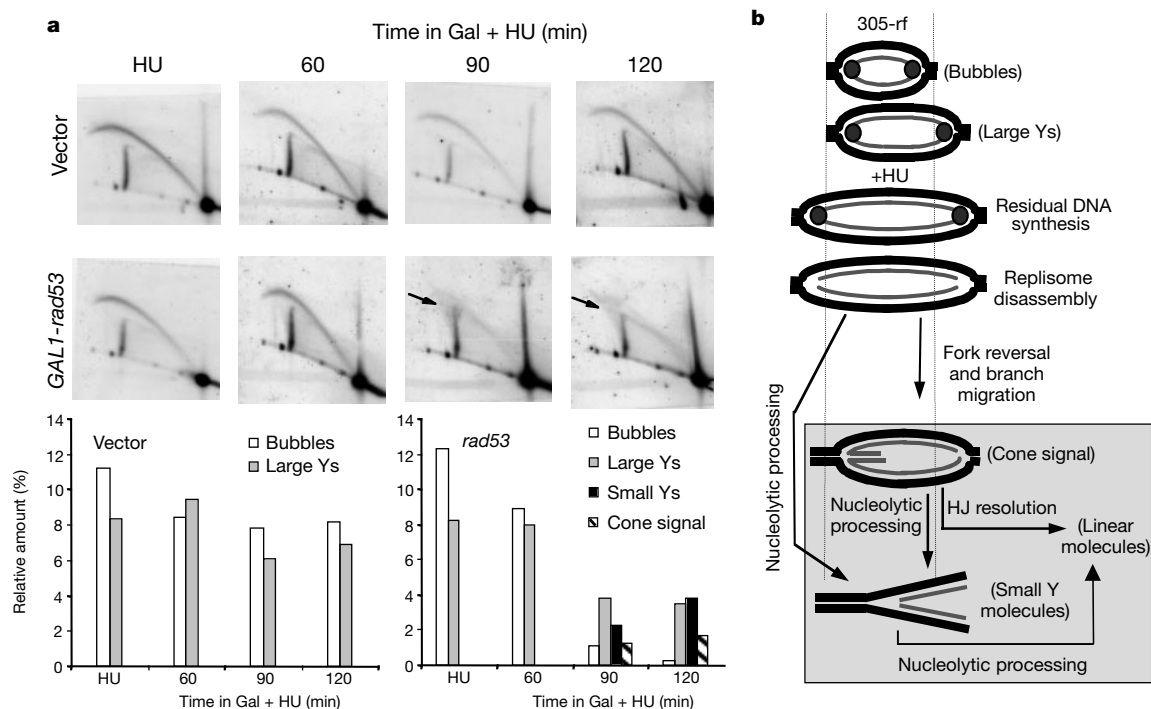


Figure 4 Proper maintenance of hydroxyurea-arrested replication forks requires Rad53 activity. **a**, CY2572 (vector) and CY2573 (*GAL1-rad53-D339A*) cells were grown in YP medium containing 2% raffinose at 28 °C, pre-synchronized by α -factor treatment and then released in fresh media containing 0.2 M hydroxyurea (HU) plus 2% raffinose to prevent the expression of the *rad53* mutant form. After 3.5 h in the presence of the drug (hydroxyurea sample), galactose (Gal) was added at 2% final concentration. DNA samples were prepared at the indicated times, and 10- μ g aliquots were cut with 100 units of *EcoRI*. After two-dimensional gel electrophoresis the blots were probed with the 3.7-kb

BamHI-EcoRI fragment spanning the *ARS305* origin. Arrows indicate the cone-shaped signal. Quantification analysis is also presented. **b**, Schematic representation of a hydroxyurea-induced replication fork arrest in *rad53* mutant cells. Dashed lines delimit the restriction fragment containing *ARS305* visualized by two-dimensional-gel analysis. The black ball represents the replisome. In *rad53* cells, stalled and unprotected replication forks (see text for details) can branch migrate on either side of the bubble and/or be processed through nucleolytic events to generate the abnormal DNA structures represented in the grey panel. HJ, Holliday junctions.

the inability of these cells to recover from the hydroxyurea treatment. We did not detect abnormal DNA structures at early replicating origins in *rad53* cells grown under normal conditions (Fig. 3c and data not shown), and therefore we must assume that hydroxyurea treatment greatly amplifies the presence of these abnormal intermediates. Although our approach may not be sensitive enough to detect a small amount of these structures, it is also possible that replication forks in the 305-rf in *rad53* cells grown under normal conditions will never collapse, but rather that this event is restricted to natural pause sites in the genome^{25,26} or sites where the forks encounter a damaged template. From this perspective, we propose that the checkpoint response directly modulates the stability of replicating chromosomes, thus contributing to the prevention of genomic rearrangements, which are the most prominent hallmarks of cancer susceptibility in multicellular organisms. □

Methods

We used the following strains: W303-1A²⁷ (*MATa ade2-1, trp1-1, leu2-3, 112 hys3-11, 15 ura3, can1-100*) and its isogenic derivatives CY2034 (*rad53-K227A-KanMX4*), CY387 (*pri1-M4*), CY2059 (*pri2-1*) and CY2061 (*cdc17-1*)^{6,28}. Strains CY2572 (vector) and CY2573 (*GAL1rad53*) were constructed by integrating in the W303-1A strain, respectively, the YIplac128 (*LEU2*) vector plasmid or its pCH12 derivative⁶, containing the *EcoRI* fragment carrying the *rad53-D339A* mutant allele under the control of the *GAL1* promoter. Strains CY3278 (*mec1-td*) and CY3281 (*mec1-td, rad53-K227A*) are isogenic to W303 and were constructed by replacing the wild-type copy of *MEC1* with the *mec1-Δdegron* allele as already described²⁹.

Yeast protein extracts prepared by the TCA extraction method⁶ were resolved by 10% SDS-PAGE, and the phosphorylation state of the Rad53 polypeptide was analysed by western blotting using anti-Rad53 antibodies (provided by C. Santocane and J. Diffley). DNA samples to be analysed with the neutral-neutral two-dimensional electrophoresis technique were prepared and analysed essentially as described¹⁷: first-dimension gels were 0.35% agarose and second-dimension gels were 0.9% agarose. Replication intermediates were quantified as already described²⁶, by calculating the percentage of the specific replication-intermediate signals relatively to the monomer spot. FACS analysis was performed as described⁶.

Received 10 May; accepted 19 June 2001.

- Elledge, S. J. Cell cycle checkpoints: preventing an identity crisis. *Science* **274**, 1664–1672 (1996).
- Weinert, T. DNA damage checkpoints update: getting molecular. *Curr. Opin. Genet. Dev.* **8**, 185–193 (1998).
- Lowndes, N. F. & Murguia, J. R. Sensing and responding to DNA damage. *Curr. Opin. Genet. Dev.* **10**, 17–25 (2000).
- Foiani, M. *et al.* DNA damage checkpoints and DNA replication controls in *Saccharomyces cerevisiae*. *Mutat. Res.* **21**, 286–294 (2000).
- Weinert, T. Yeast checkpoints controls and relevance to cancer. *Cancer Surv.* **29**, 109–132 (1997).
- Pelliccioli, A. *et al.* Activation of Rad53 kinase in response to DNA damage and its effect in modulating phosphorylation of the lagging strand DNA polymerase. *EMBO J.* **18**, 6561–6572 (1999).
- Paulovich, A. G. & Hartwell, L. H. A checkpoint regulates the rate of progression through S-phase in *S. cerevisiae* in response to DNA damage. *Cell* **82**, 841–847 (1995).
- Santocane, C. & Diffley, J. F. X. A *Mec1*- and Rad53-dependent checkpoint controls late-firing origins of DNA replication. *Nature* **395**, 615–618 (1998).
- Shirahige, K. *et al.* Regulation of DNA-replication origins during cell-cycle progression. *Nature* **395**, 618–621 (1998).
- Desany, B. A., Alcasabas, A. A., Bachant, J. B. & Elledge, S. J. Recovery from DNA replication stress is the essential function of the S-phase checkpoint pathway. *Genes Dev.* **12**, 2956–2970 (1998).
- Brush, G., Morrow, D. M., Heiter, P. & Kelly, T. J. The ATM homolog *Mec1* is required for phosphorylation of replication protein A in yeast. *Proc. Natl Acad. Sci. USA* **93**, 15075–15080 (1996).
- Liberti, G. *et al.* Srs2 DNA helicase is involved in checkpoint response and its regulation requires a functional *Mec1*-dependent pathway and CDK1 activity. *EMBO J.* **19**, 5027–5038 (2000).
- Bashkirov, V. I. *et al.* DNA repair protein Rad55 is a terminal substrate of the DNA damage checkpoints. *Mol. Cell. Biol.* **20**, 4393–4404 (2000).
- Brewer, B. J. & Fangman, W. L. The localization of replication origins on ARS plasmids in *S. cerevisiae*. *Cell* **51**, 463–471 (1987).
- Zhao, X., Muller, E. G. & Rothstein, R. A suppressor of two essential checkpoint genes identifies a novel protein that negatively affects dNTP pools. *Mol. Cell* **2**, 329–340 (1998).
- Newlon, C. S. *et al.* Analysis of replication origin function on chromosome III of *Saccharomyces cerevisiae*. *Cold Spring Harb. Symp. Quant. Biol.* **58**, 415–423 (1993).
- Friedman, K. L. & Brewer, B. J. Analysis of replication intermediates by two-dimensional agarose gel electrophoresis. *Methods Enzymol.* **262**, 613–627 (1995).
- Lindsay, H. D. *et al.* S-phase-specific activation of Cds1 kinase defines a subpathway of the checkpoint response in *Schizosaccharomyces pombe*. *Genes Dev.* **12**, 382–395 (1998).
- Edwards, R. J., Bentley, N. J. & Carr, A. M. A Rad3–Rad26 complex response to DNA damage independently of other checkpoint proteins. *Nature Cell Biol.* **1**, 393–398 (1999).
- Dimitrova, D. S. & Gilbert, D. M. Temporally coordinated assembly and disassembly of replication factories in the absence of DNA synthesis. *Nature Cell Biol.* **2**, 686–694 (2000).
- Seigneur, M., Bidnenko, V., Ehrlich, S. D. & Michel, B. RuvAB acts at arrested replication forks. *Cell* **95**, 419–430 (1998).
- Rothstein, R., Michel, B. & Gangloff, S. Replication fork pausing and recombination or “gimme a break”. *Genes Dev.* **14**, 1–10 (2000).
- Martin-Parras, L., Hernandez, P., Martinez-Robles, M. L. & Schwartzman, J. B. Initiation of DNA

replication in ColE1 plasmids containing multiple potential origins of replication. *J. Biol. Chem.* **267**, 22496–22505 (1992).

- Kalejta, R. F. & Hamlin, J. L. Composite patterns in neutral/neutral two-dimensional gels demonstrate inefficient replication origin usage. *Mol. Cell. Biol.* **16**, 4915–4922 (1996).
- Deshpande, A. M. & Newlon, C. S. DNA replication fork pause sites dependent on transcription. *Science* **272**, 1030–1033 (1996).
- Ivessa, A. S., Zhou, J. Q. & Zakian, V. A. The *Saccharomyces* Pif1p DNA helicase and the highly related Rrm3p have opposite effects on replication fork progression in ribosomal DNA. *Cell* **100**, 479–489 (2000).
- Thomas, B. J. & Rothstein, R. Elevated recombination rates in transcriptionally active DNA. *Cell* **56**, 619–629 (1989).
- Marini, F. *et al.* Role for DNA primase in coupling DNA replication to DNA damage response. *EMBO J.* **16**, 639–650 (1997).
- Pelliccioli, A., Lee, S. E., Lucca, C., Foiani, M. & Haber, J. E. Regulation of *Saccharomyces* Rad53 checkpoint kinase during adaptation from DNA damage-induced G2/M arrest. *Mol. Cell* **7**, 293–300 (2001).

Supplementary information is available from Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank S. Alberti, L. Fabiani, A. Gambetta, S. Pintus and J. Theis for technical advice and support. We also thank A. Carr, J. Diffley, J. Haber, R. Rothstein, J. Sogo and all the members of our laboratory for helpful discussions. This work was supported by Associazione Italiana per la Ricerca sul Cancro and partially by grants from Telethon–Italy, Cofinanziamento MURST–Università di Milano, MURST (5%) Biomolecole per la Salute Umana, and CNR Target Project on Biotechnology, by a EU TMR contract, and by a NIH grant to C.S.N.

Correspondence and requests for materials should be addressed to M.F. (e-mail: foiani@ifom-firc.it).

Rb targets histone H3 methylation and HP1 to promoters

Soren J. Nielsen^{*†}, Robert Schneider^{*†}, Uta-Maria Bauer^{*}, Andrew J. Bannister^{*}, Ashby Morrison[‡], Donal O'Carroll[§], Ron Firestein^{||}, Michael Cleary^{||}, Thomas Jenuwein[§], Rafael E. Herrera[‡] & Tony Kouzarides^{*}

^{*} Wellcome/CRC Institute and Department of Pathology, Tennis Court Road, Cambridge CB2 1QR, UK

[‡] Baylor College of Medicine, Department of Molecular and Cellular Biology, The Breast Center, 1 Baylor Plaza, Houston, Texas 77030, USA

[§] Research Institute of Molecular Pathology (IMP), The Vienna Biocenter, Dr. Bohrgasse 7, A-1030 Vienna, Austria

^{||} Department of Pathology, Stanford University Medical Center, Stanford, California 94305, USA

[†] These authors have contributed equally to this work.

In eukaryotic cells the histone methylase SUV39H1 and the methyl-lysine binding protein HP1 functionally interact to repress transcription at heterochromatic sites¹. Lysine 9 of histone H3 is methylated by SUV39H1 (ref. 2), creating a binding site for the chromo domain of HP1 (refs 3, 4). Here we show that SUV39H1 and HP1 are both involved in the repressive functions of the retinoblastoma (Rb) protein. Rb associates with SUV39H1 and HP1 *in vivo* by means of its pocket domain. SUV39H1 cooperates with Rb to repress the cyclin E promoter, and in fibroblasts that are disrupted for SUV39, the activity of the cyclin E and cyclin A2 genes are specifically elevated. Chromatin immunoprecipitations show that Rb is necessary to direct methylation of histone H3, and is necessary for binding of HP1 to the cyclin E promoter. These results indicate that the SUV39H1–HP1 complex is not only involved in heterochromatic silencing but also has a role in repression of euchromatic genes by Rb and perhaps other co-repressor proteins.

The Rb protein functions as a repressor, at least partly, through the recruitment of histone deacetylase activity^{5–7}. We considered whether histone methylation might also be involved in Rb-mediated

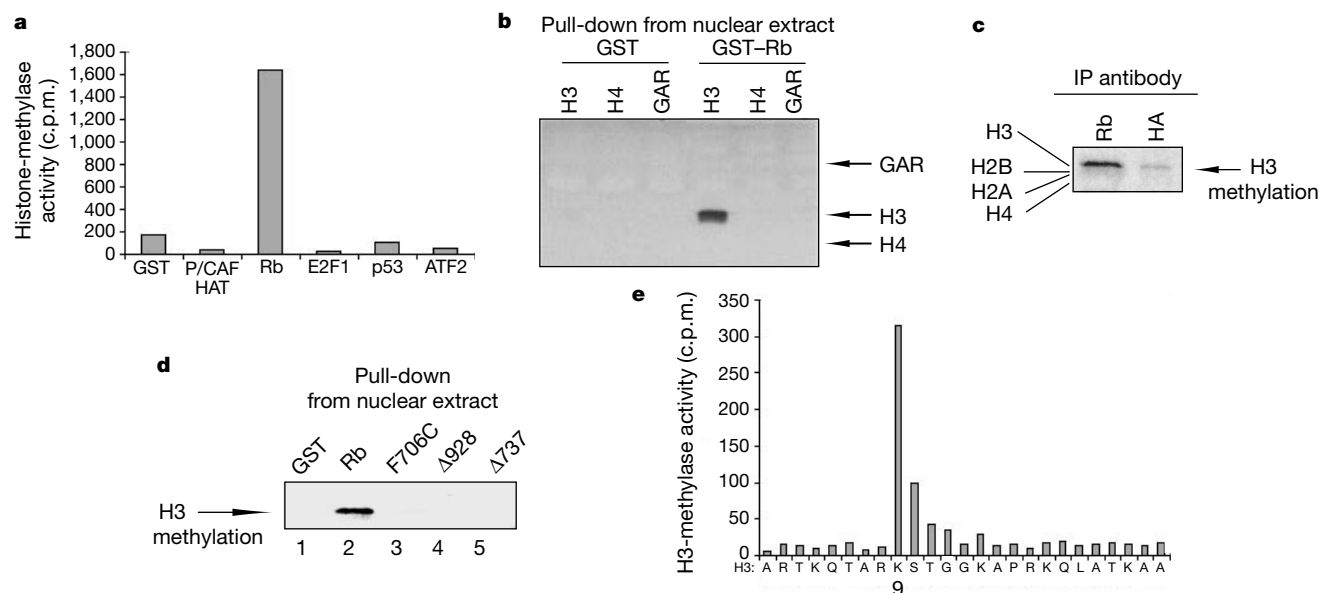


Figure 1 Rb interacts with methylase activity specific for H3 Lys 9. **a**, GST fusion proteins (2 μ g) were used to purify histone methylase activity from 500 μ g of HeLa nuclear extract. c.p.m., counts per minute. **b**, Rb-associated activity methylates H3 but not H4 or the arginine-methylase substrate GAR. **c**, Endogenous Rb associates with H3-specific methylase activity. HeLa nuclear extract was immunoprecipitated using a Rb-specific

antibody or a control antibody (HA, 2 μ g each). Immunoprecipitates were tested for associated methylase activity. **d**, Mutations in the Rb pocket disrupt the Rb-methylase interaction as they do not associate with methylase activity. **e**, H3 labelled by Rb-associated methylase was sequenced. Fractions corresponding to each amino-acid cycle were collected and analysed by scintillation counting.

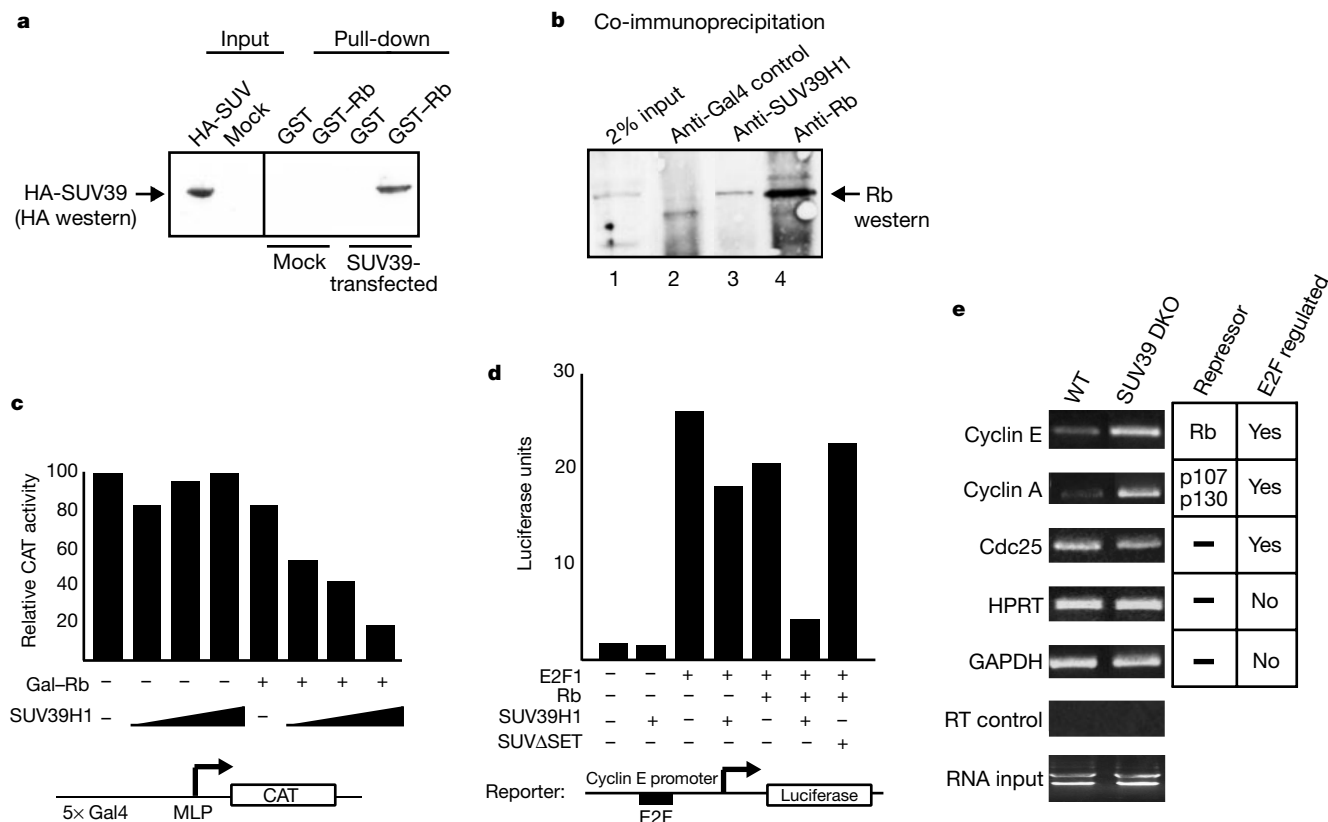


Figure 2 SUV39H1 and Rb interact and regulate transcription. **a**, Rb purifies SUV39H1 from cells. HEK-293 cells were transfected with a HA-SUV39H1 or an empty expression vector (Mock). Extracts were incubated with GST or GST-Rb (4 μ g) and washed. Bound SUV39H1 was western blotted using an anti-HA antibody. **b**, SUV39H1 and Rb form a complex *in vivo*. U2OS cell nuclear extract was immunoprecipitated with antibodies (10 μ g) against SUV39H1, Rb or Gal4-DBD. Immunoprecipitates were western blotted with a Rb antibody. **c**, U2OS cells transfected with a Gal4-driven CAT reporter (0.33 μ g) under the control of the major late promoter (MLP) together with an expression vector for Gal4-Rb or Gal4 alone (0.66 μ g) plus increasing amounts of a SUV39H1 expression

vector (0.13–2 μ g). **d**, HeLa cells were transfected with a reporter containing the cyclin E promoter driving luciferase (5 μ g) together with combinations of expression vectors for E2F1 (2 μ g), Rb (2 μ g), SUV39H1 (0.1 μ g) and SUV39H1 Δ SET (0.1 μ g). **e**, RNA from wild-type and SUV39H1 and SUV39H2 double-knockout (DKO) mice was isolated. Equal amounts of RNA (RNA input) were analysed by RT-PCR (25 cycles) for the expression of cyclin E, cyclin A2, Cdc25C, GAPDH and HPRT. The RT control lanes represent RT-PCR reactions in the absence of reverse transcription. The RNA used here was from cells of female mice, but identical results were obtained using RNA from cells of male mice (data not shown).

repression, as the SUV39H1 methylase has repressive potential⁸. To establish whether Rb can associate with histone-methylase activity, a glutathione *S*-transferase (GST)–Rb fusion was incubated with nuclear extract, and any bound methylase activity was assayed on bulk histones as a substrate. Figure 1a shows that GST–Rb (but not GST alone) can purify histone-methylase activity, whereas GST fusions to transcriptional activators such as P/CAF, E2F1, p53 and ATF2 do not. The Rb-associated methylase activity is specific for histone H3 and does not recognize the GAR substrate for arginine methylases (Fig. 1b).

An antibody directed against Rb can precipitate histone-methylase activity that is specific for histone H3 (Fig. 1c). This methylase binds the pocket domain of Rb because tumour-derived mutations in the pocket (F706C), or truncations of the pocket (Δ 928 and Δ 737), abolish binding to the methylase (Fig. 1d). The Rb-associated methylase has specificity for Lys 9 of histone H3, as shown by Edman degradation of radioactively methylated histone H3 (Fig. 1e).

The SUV39H1 protein possesses lysine methylase activity, which resides within its conserved SET domain². As this enzyme has specificity for Lys 9 of histone H3 we investigated whether SUV39H1 could be the methylase associated with Rb. Figure 2a shows that a GST–Rb fusion can bind to transfected, haemagglutinin (HA)-tagged SUV39H1. Endogenous Rb also associates with endogenous SUV39H1, as shown by a co-immunoprecipitation analysis (Fig. 2b). As DNA is present in these reactions as a low-level contaminant, it remains a formal possibility that the interaction is facilitated by DNA; however, as histones are not co-immunoprecipitated this possibility seems unlikely (data not shown).

We next investigated whether SUV39H1 could act as co-repressor with Rb. Figure 2c shows that SUV39H1 represses the activity of a promoter bearing *GAL4* sites in a concentration-dependent manner *in vivo*, but only when Gal4–Rb is present at the promoter. The co-repressor functions of SUV39H1 can also be seen on the cyclin E promoter, a natural target for Rb-mediated repression^{9,10}. This

promoter can be stimulated by E2F (Fig. 2d, columns 1 and 3) and is not affected by SUV39H1 alone (columns 1 and 2). Under limiting conditions, where Rb represses E2F activity slightly (column 5), the SUV39H1 enzyme can further repress E2F activity in cooperation with Rb (column 6). When the methylase domain of SUV39H1 is removed the resulting SUV39H1 Δ SET is unable to mediate repression (column 7). These results suggest that SUV39H1 uses its methylase activity to repress the cyclin E promoter when it is targeted there by Rb.

The repressive functions of SUV39H1 were verified using RNA isolated from fibroblasts that lacked both SUV39H1 and the closely related methylase SUV39H2 (double-knockout cells). Reverse transcription followed by polymerase chain reaction (RT-PCR) was used to show that cyclin E messenger RNA levels are elevated in the double-knockout cells compared with wild-type cells (Fig. 2e). mRNA of cyclin A2, a gene repressed by the Rb-related pocket proteins p107 and p130 (ref. 9), is also upregulated in double-knockout cells. In contrast, the expression of *Cdc25C*, a gene regulated by E2F but not repressed by Rb, p107, or p130 (ref. 9), is unchanged in double-knockout cells. The expression of two unrelated house-keeping genes, *GAPDH* and *HPRT*, is also unchanged. Collectively, these results support the conclusion that the SUV39H1 methylase is a specific repressor of genes regulated by the Rb-pocket family.

SUV39H1 is known to be in a complex with the HP1 protein¹¹. Recently, HP1 function has been placed downstream of SUV39H1 histone methylation, as HP1 recognizes specifically, and binds to, histone H3 methylated at Lys 9 (refs 1, 4, 5). This mechanistic link prompted us to investigate the role of HP1 in Rb/SUV39H1-mediated repression. Rb and HP1 can interact in a two-hybrid screen in yeast, and it has been shown that there is an LXCXE motif (X is any amino acid) in HP1 (ref. 12). We therefore asked if HP1 binds to Rb in mammalian cells. Figure 3a shows that a GST–HP1 fusion can bind Rb that is present in nuclear extracts; Rb and

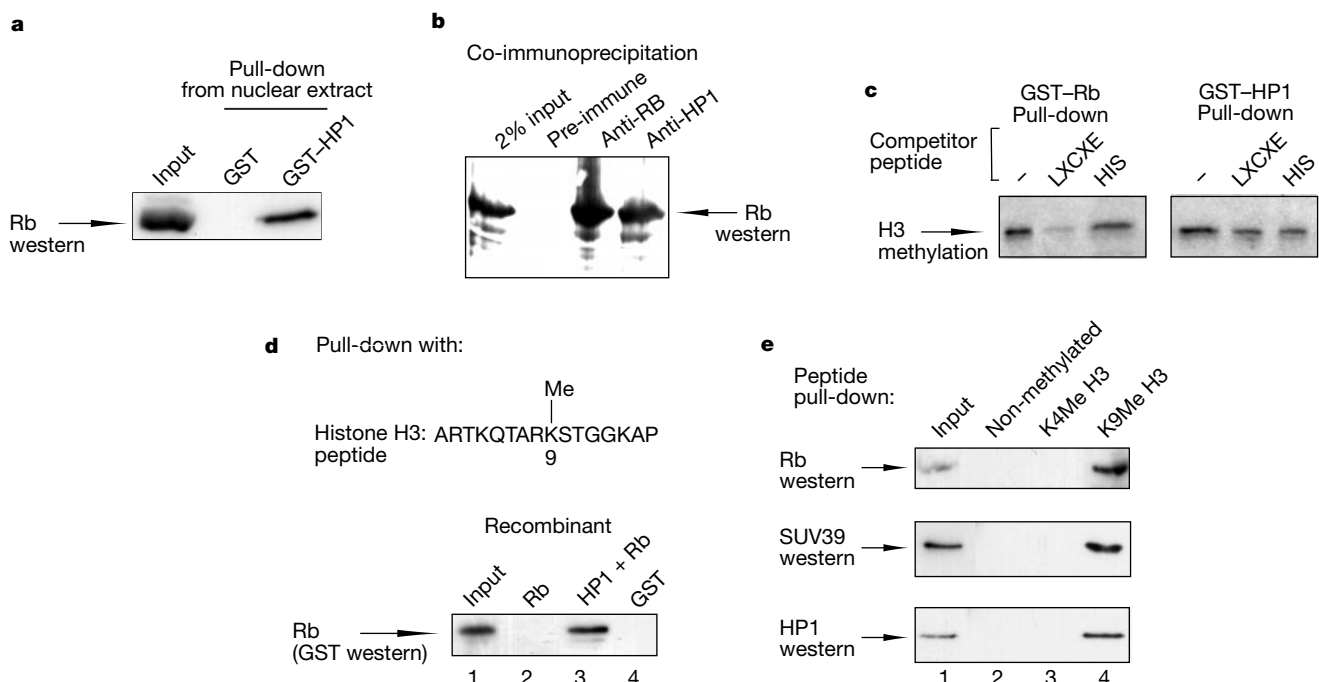


Figure 3 HP1 interacts with Rb in an LXCXE-dependent fashion. **a**, GST fusion proteins (4 μ g) were tested for binding to Rb from HeLa nuclear extract. **b**, Endogenous HP1 and Rb interact. U2OS nuclear extract was immunoprecipitated with anti-HP1 serum, anti-Rb antibody, or a pre-immune serum, and western blotted for Rb. **c**, GST–Rb and GST–HP1 (2 μ g) were tested for binding to H3-specific methylase from HeLa nuclear extract in the presence of an LXCXE or control (HIS) peptide. **d**, HP1 recruits Rb to H3 methylated at

Lys 9. A H3 peptide methylated at Lys 9 (2 μ g) immobilized on beads was incubated with recombinant GST–Rb (0.5 μ g) in the absence and presence of His-HP1 (1 μ g). Rb binding was detected by western blotting. **e**, Rb, HP1 and SUV39H1 bind specifically to H3 methylated at Lys 9 (K9Me H3). Differentially modified H3 peptides were tested for binding of endogenous HP1 and SUV39H1 and transfected Rb from cell extracts. Input lanes contain 2% of total extract.

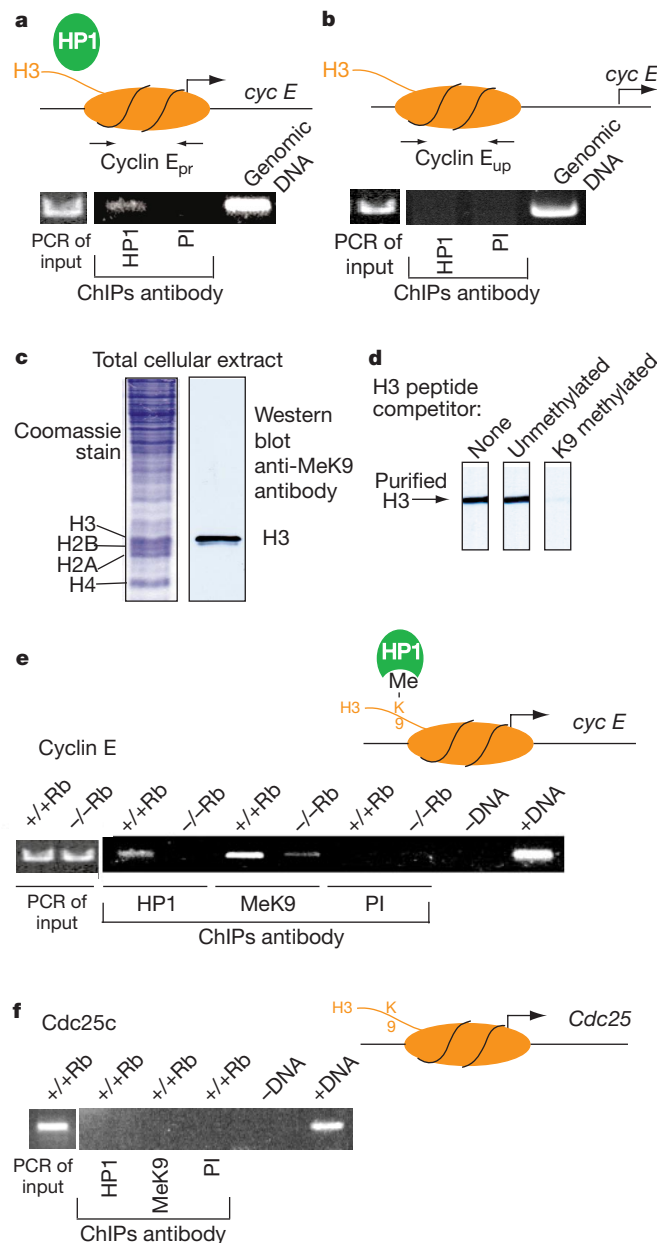


Figure 4 Rb is required for HP1 promoter recruitment. **a, b**, A specific nucleosome within the cyclin E promoter associates with HP1. Chromatin immunoprecipitations (ChIPs) from wild-type MEFs were performed with HP1 antiserum, or a pre-immune (PI) control, and the purification of cyclin E promoter fragments (centred at +1 for cyclin E_{pr} in **a**; -550 for cyclin E_{up} in **b**) was analysed by quantitative PCR. **c, d**, Characterization of an anti-H3 methyl Lys 9 antibody (anti-MeK9). Total cellular extract was visualized by Coomassie blue staining or probed with anti-MeK9. H3 peptide competition showed that the antibody was only effectively competed against by a histone H3 peptide when methylated at Lys 9. **e**, Chromatin immunoprecipitations show Rb dependence for HP1 recruitment to the cyclin E promoter. Crosslinked chromatin from Rb^{+/+} and Rb^{-/-} cells was immunoprecipitated with HP1 antiserum, anti-MeK9 antibody or pre-immune control. Equal abundance of cyclin E promoter sequence in Rb^{+/+} and Rb^{-/-} nucleosomal preparations was determined by PCR from the input chromatin. **f**, The *Cdc25C* gene does not associate with HP1 or histone H3 methylated at Lys 9. Chromatin immunoprecipitations were performed as in **e**.

HP1 can associate *in vivo*, as determined by co-immunoprecipitation analysis (Fig. 3b). An LXCXE motif peptide can compete for the binding of histone H3 methylase activity to Rb, but does not affect the binding of H3 methylase activity to HP1 (Fig. 3c), which is consistent with the finding that the methylase activity is associated with the Rb pocket.

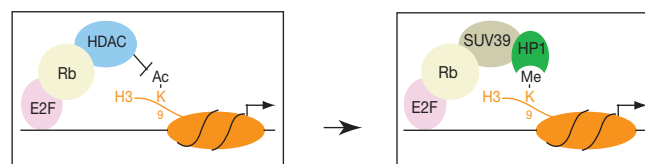


Figure 5 Role of the SUV39H1-HP1 complex in the transcriptional co-repressor function of Rb. Deacetylation of histone H3 at Lys 9 by Rb-associated deacetylase activity (HDAC) might be required as a preceding step to SUV39H1-mediated methylation.

We next assessed whether HP1 can recognize methylated lysines while associated with Rb. To address this, a histone H3 peptide methylated at Lys 9 was used as an affinity resin. Recombinant Rb does not bind to this methylated peptide (Fig. 3d, lane 2), but it can do so efficiently in the presence of recombinant HP1 (Fig. 3d, lane 3). This result confirms that HP1 can bind directly to Rb and that it can recognize Rb and methylated lysine simultaneously. A similar experiment was attempted using nuclear extracts as the source of protein. Figure 3e shows that the H3 peptide methylated at Lys 9 (but not unmethylated or Lys-4-methylated peptide) binds to HP1, SUV39H1 and Rb, as detected by western blotting.

The above results suggest that an Rb-regulated promoter such as cyclin E should be associated with HP1. To test this we performed chromatin immunoprecipitation analysis of the cyclin E promoter. Figure 4a shows that a nucleosome encompassing the cyclin E initiation site (cyclin E_{pr}) that is known to be deacetylated (A.M. and R.H., personal communication) is associated with HP1 in fibroblast cells of mouse embryos. In contrast an upstream nucleosome (cyclin E_{up}) does not associate with HP1 (Fig. 4b). As the cyclin E_{pr} nucleosome binds HP1, we next addressed whether this nucleosome contains histone H3 that is methylated at Lys 9. To test this we raised an antibody that recognizes histone H3 when methylated at Lys 9 (Fig. 4c, d). Figure 4e shows that in Rb^{+/+} cells the cyclin E_{pr} nucleosome contains methylated histone H3 and is associated with HP1. However, in Rb^{-/-} cells histone H3 methylation and HP1 binding is significantly reduced. A nucleosome encompassing the *Cdc25C* initiation site¹³ (a gene that is not repressed by Rb¹⁰) contains unmethylated histone H3 and does not bind HP1 (Fig. 4f). Thus, in the presence of Rb, methylase activity and HP1 are targeted to the cyclin E promoter.

Collectively, the results presented here implicate each of the components of the SUV39H1-HP1 complex in the repression functions of the Rb protein. In this model (Fig. 5) Rb brings to the promoter the SUV39H1 enzyme (and possibly other members of this family) to methylate Lys 9 of histone H3 and provide a binding site for HP1. Methylation by SUV39H1 cannot take place on an already acetylated lysine². Thus the deacetylase activity associated with Rb⁵⁻⁷ may be a necessary preceding step to SUV39H1-mediated methylation. The precise function of HP1 in repression is unclear. HP1 may protect the methyl group on Lys 9 from attack from potential demethylases, it may bring in other repressive functions, or it may enhance the stability of the Rb-associated repressor complex.

HP1 is found associated with a number of transcriptional repressors¹, suggesting that it may have a role in repressing many other promoters. Thus, the results presented here extend the role of SUV39H1 and HP1 beyond heterochromatic gene silencing to a more general, genome-wide function in repressing gene transcription.

Methods

Cell culture, transfections and transcription assays

Cells (U20S and HEK-293) were transfected using the calcium phosphate technique or FuGene (Roche) according to the manufacturer's instructions. Twenty-four hours after

transfection, cells were collected and processed for CAT or luciferase activity using standard techniques¹⁴.

GST pull downs and immunoprecipitations

GST-Rb (wild type and mutants¹⁵) and other GST fusion proteins were expressed and purified from *Escherichia coli* XA90 (ref. 16). GST fusion proteins that were immobilized on glutathione-sepharose (Pharmacia), or H3-derived peptides³ bound to Sulfolink beads (Pierce), were incubated with extract in IPH buffer¹⁶. Complexes were washed four times in IPH buffer before processing for methylase assays or western blotting. Antibodies against HA (12CA5, Roche), Gal4-DBD (DNA-binding domain; sc-510, Santa Cruz), SUV39H1 (M. Cleary), Rb (G3-245; XZ55, PharMingen) or HP1 (ref. 3) were used. For immunoprecipitations antibodies were incubated with HeLa nuclear extract (Cell Culture Center) or U2OS nuclear extract in IPH buffer at 4 °C (ref. 17). After 2 h a 50:50 mixture of protein A/G-sepharose beads (Pharmacia) was added. To avoid the possibility that DNA mediates the interaction between SUV39H1 and Rb, the immunoprecipitations were probed for the presence of histones with negative results.

Histone methylase assays and protein sequencing

Precipitations from pull downs or immunoprecipitations were incubated with 20 µg histones (Sigma) and 1 µl [3H-Me]-S-adenosyl methionine (NEN, 80 Ci mmol⁻¹) in PBS at 30 °C for 1 h. Assays were analysed by SDS-PAGE followed by western blotting and autoradiography or were spotted onto P-81 cationic exchange paper (Whatman), washed in carbonate buffer and quantified by scintillation counting³. For amino-terminal sequencing, radiolabelled H3 was blotted to polyvinylidene fluoride and sequenced by Edman degradation (Protein Sequencing Facility, University of Cambridge). We counted fractions for the presence of tritium.

RNA purification and RT-PCR analysis

Total RNA (0.5 µg) was isolated from W12 (wild type) and D3 (SUV39H1 and SUV39H2 double knockout; D.O. and T.J., unpublished observations) female mouse cells, and was used for quantitative RT-PCR, following the Qiagen One Step protocol, for 20, 25 and 30 PCR cycles.

Antibody generation

Rabbits were immunized with a H3 N-terminal lysine-methylated peptide corresponding to amino acids 1–16. Immunoreactive serum was applied to a H3 Lys-9-methylated peptide column to affinity purify specific antibodies, as the antiserum crossreacted with H3 methylated at Lys 4.

Chromatin immunoprecipitation

Chromatin immunoprecipitations were performed using HeLa cells and MEF cells essentially as described^{18,19}. Immunoprecipitates were analysed for the presence of cyclin E or *Cdc25C* promoter fragments by PCR using primers specific for single nucleosomes. PCR reactions were repeated exhaustively using varying cycle numbers and different amounts of templates to ensure that results were within the linear range of the PCR.

Received 6 April; accepted 6 July 2001.

- Jones, D. O., Cowell, I. G. & Singh, P. B. Mammalian chromodomain proteins: their role in genome organisation and expression. *BioEssays* **22**, 124–137 (2000).
- Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593–599 (2000).
- Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromodomain. *Nature* **410**, 120–124 (2001).
- Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 120–124 (2001).
- Brehm, A. *et al.* Retinoblastoma protein recruits histone deacetylase to repress transcription. *Nature* **391**, 597–601 (1998).
- Magnaghi-Jaulin, L. *et al.* Retinoblastoma protein represses transcription by recruiting a histone deacetylase. *Nature* **391**, 601–605 (1998).
- Luo, R. X., Postigo, A. A. & Dean, D. C. Rb interacts with histone deacetylase to repress transcription. *Cell* **92**, 463–473 (1998).
- Firestein, R., Cui, X., Huie, P. & Cleary, M. L. Set domain-dependent regulation of transcriptional silencing and growth control by SUV39H1, a mammalian ortholog of *Drosophila* Su(var)3-9. *Mol. Cell. Biol.* **20**, 4900–4909 (2000).
- Hurford, R. K. Jr, Cobrinik, D., Lee, M. & Dyson, N. pRB and p107/p130 are required for regulated expression of different sets of E2F responsive genes. *Genes Dev.* **11**, 1447–1463 (1997).
- Herrera, R. E. *et al.* Altered cell cycle kinetics, gene expression and G1 restriction point regulation in Rb deficient fibroblasts. *Mol. Cell. Biol.* **16**, 2402–2407 (1996).
- Aagaard, L. *et al.* Functional mammalian homologues of the *Drosophila* PEV-modifier Su(var)3-9 encode centromere-associated proteins which complex with the heterochromatin component M31. *EMBO J.* **18**, 1923–1938 (1999).
- Williams, L. & Grafi, G. The retinoblastoma protein—a bridge to heterochromatin. *Trends Plant Sci.* **5**, 239–240 (2000).
- Korner, K. & Muller, R. J. *In vivo* structure of the cell cycle-regulated human *cdc25C* promoter. *Biol. Chem.* **275**, 18676–18681 (2000).
- Hagemeier, C., Cook, A. & Kouzarides, T. The retinoblastoma protein binds E2F residues required for activation *in vivo* and TBP binding *in vitro*. *Nucleic Acids Res.* **21**, 4998–5004 (1993).
- Qin, X. Q., Chittenden, T., Livingston, D. M. & Kaelin, W. G. Jr Identification of a growth suppression domain within the retinoblastoma gene product. *Genes Dev.* **6**, 953–964 (1992).
- Bannister, A. J. & Kouzarides, T. The CBP co-activator is a histone acetyltransferase. *Nature* **384**, 641–643 (1996).

- Nielsen, A. L. *et al.* Interaction with members of the heterochromatin protein 1 (HP1) family and histone deacetylation are differentially involved in transcriptional silencing by members of the TIF1 family. *EMBO J.* **18**, 6385–6395 (1999).
- Orlando, V., Strutt, H. & Paro, R. Analysis of chromatin structure by *in vivo* formaldehyde cross-linking. *Methods* **11**, 205–214 (1997).
- Dedon, P. C., Soult, J. A., Allis, C. D. & Gorovsky, M. A. A simplified formaldehyde fixation and immunoprecipitation technique for studying protein–DNA interactions. *Anal. Biochem.* **197**, 83–90 (1991).

Acknowledgements

We thank M. Weldon for Edman degradation of labelled proteins; H. Herschman for providing GAR; R. Laskey for the anti-HP1 antibody; and A. Cook for technical assistance. S.J.N. and A.J.B. were funded by a grant from the Cancer Research Campaign, R.S. by an EC grant and an EMBO fellowship, and U.M.B. by an HFSP grant.

Correspondence and requests for materials should be addressed to T.K. (e-mail: tk106@mole.bio.cam.ac.uk).

correction

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium

Nature **409**, 860–921 (2001).

We have identified several items requiring correction or clarification in our paper on the sequencing of the human genome.

● Six additional authors should have been included: Pieter de Jong, Joseph J. Catanese, and Kazutoyo Osoegawa (Department of Cancer Genetics, Roswell Park Cancer Institute, Buffalo, New York 14263, USA; present address: Children's Hospital Oakland Research Institute, 747 52nd street Oakland, California 94609, USA) and Hiroaki Shizuya, Sangdun Choi and Yu-Juin Chen (Division of Biology, California Institute of Technology, Pasadena, California 91125, USA). These investigators and their laboratories constructed the high-quality BAC libraries that were crucial in sequencing the genome, as described in Table 1. These libraries were not previously published. We apologize to our colleagues for this omission.

● The Supplementary Information on *Nature's* website has been revised. Changes to the original Supplementary Information are available in the Supplementary Information to this Correction. We have added 7 additional investigators to the full list of authors. We have also added 79 additional references, citing previously published sequences that were included in the draft genome sequence.

● Table 27 reported 18 instances of apparently novel paralogues of genes encoding drug targets. We have carefully reviewed these 18 cases and found that two are incorrect: a paralogue of an insulin-like growth factor-1 receptor gene and a paralogue of the calcitonin-related polypeptide alpha gene. In both cases, we had incorrectly recorded the chromosomal location sequence of the known gene, thereby erroneously giving rise to an apparent paralogue (the first instance was identified by J. Englebrecht and C. Kristensen (personal communication)). Of the 16 remaining apparent paralogues, two (calcium channel paralogue IGI_M1_ctg17137_10 and heparan N-deacetylase/N-sulphotransferase paralogue IGI_M1_ctg13263_18) have so far been confirmed as bona fide genes^{1,2}.

● Several correspondents have written to point out that a handful of clones listed as human sequence in the HTG division of GenBank (established to house 'unfinished' sequence data) are actually mouse sequence (about two dozen out of 30,000 clones). They asked

whether these clones give rise to contamination in the human draft sequence. As noted in the paper, we used computer programs to identify and eliminate instances of such contamination (with mouse sequence, vector sequence, and so on) before assembling the draft genome sequence. In reviewing the work, we identified one mouse clone that slipped through the filter. This clone has been eliminated in subsequent assemblies (<http://genome.cse.ucsc.edu/>). Because the draft sequence remains an imperfect partial product, we welcome additional comments that could help in improving it.

● The discussion of possible horizontal gene transfer from bacterial genomes to vertebrate genomes has provoked considerable discussion^{3–5}. We reported 113 instances of human genes that had reasonably close homologues in bacteria, but either had no homologue or only a weaker homologue in non-vertebrate eukaryotes for which extensive genomic sequence was available. We suggested two hypotheses to explain these data: horizontal gene transfer (HGT) from bacteria to human or gene loss in the other lineages. We had no data to distinguish between these hypotheses, although we suggested that the latter was a more “parsimonious” explanation as it involved fewer independent events. In the introduction we stated that this seemed “likely”.

Several correspondents have undertaken more comprehensive analyses and have argued that a significant proportion of the cases can be explained by gene loss^{3–5}. We agree. We believe that the two hypotheses cannot be distinguished on the basis of parsimony, because too little is known about the relative rates of HGT and gene loss in evolution. Instead, extensive sequence data from many additional organisms will be required to assess definitively the provenance of each gene.

We note that the process of HGT into the vertebrate genome from other organisms has clearly occurred on multiple occasions, as seen from the sudden arrival of many DNA transposons with strong similarities to other organisms. The most recent documented cases occurred subsequent to the eutherian radiation (see Fig. 19).

● A key reference concerning 3'-transduction by LINE elements was omitted on page 887. The sentence citing references 205 and 206 should also have cited Goodier *et al.*⁶.

● In Fig. 33, the unit on the y axis should be bp, not kb. The legend should read: “Sequence properties of segmental duplications. Distributions of length and per cent nucleotide identity are shown as a function of the number of aligned bp from the finished vs finished human genomic sequence dataset. Intrachromosomal (blue), interchromosomal (red).”

● In Fig. 41, the legend should begin: “For each of the 27 common domain families, the number of different Pfam domain types that co-occur with the family in each of the five eukaryotic proteomes. The 27 families were chosen to include the 10 most common domain families in each proteome. The data are ranked ...”

● In Table 22, the entry 81,126 should be 8,126.

● On page 898, line 31, the final phrase of the sentence (“... and the representativeness of currently ‘known’ human genes”) should be deleted. The sentence should read: “Before discussing the gene predictions for the human genome, it is useful to consider background issues, including previous estimates of the number of human genes and lessons learned from worms and flies.”

● On page 900, line 38, remove “(see above)”.

● We failed to acknowledge the crucial role of sequence editing software, which has been widely used for inspection and subsequent finishing of the sequence assemblies. The two principal programs used were CONSED⁷ and GAP4⁸. □

- Burgess, D. L. *et al.* A cluster of three novel (Ca²⁺) channel gamma subunit genes on chromosome 19q13.4: Evolution and expression profile of the gamma subunit gene family. *Genomics* **71**, 339–350 (2001).
- Aikawa, J. *et al.* Multiple isozymes of heparan sulfate/heparin GlcNAc N-deacetylase/GlcN N-sulfotransferase. Structure and activity of the fourth member, NDST4. *J. Biol. Chem.* **276**, 5876–5882 (2001).
- Salzberg, S. L. *et al.* Microbial genes in the human genome: Lateral transfer or gene loss? *Science* **292**, 1903–1906 (2001).

- Stanhope, M. J. *et al.* Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* **411**, 940–944 (2001).
- Reelofs, J. & Van Haastert, P. J. M. Genes lost during evolution. *Nature* **411**, 1013–1014 (2001).
- Goodier, J. L., Ostertag, E. M. & Kazazian, H. H. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657 (2000).
- Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **3**, 195–202 (1998).
- Staden, R., Beal, K. F. & Bonfield, J. K. The Staden package. 1998. *Methods Mol. Biol.* **132**, 115–130 (2000).

Supplementary information (with changes to the original Supplementary Information) is available on Nature's World-Wide Web site (<http://www.nature.com>).

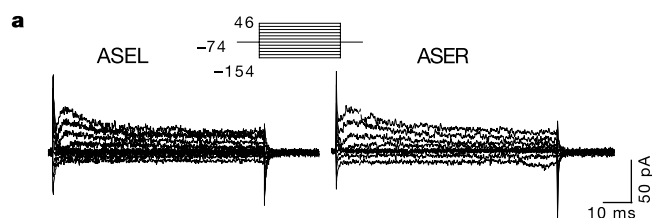
errata

The homeobox gene *lim-6* is required for distinct chemosensory representations in *C. elegans*

Jonathan T. Pierce-Shimomura, Serge Faumont, Michelle R. Gaston, Bret J. Pearson & Shawn R. Lockery

Nature **410**, 694–698 (2001).

In some United States issues of *Nature* the lines in Fig. 3a were missing or very faint. A corrected version is printed below. □

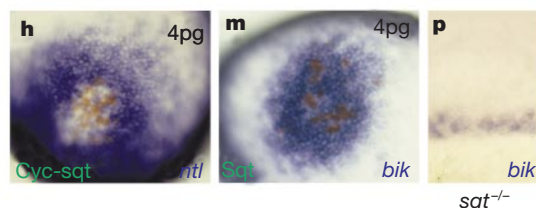


The zebrafish Nodal signal Squint functions as a morphogen

Yu Chen & Alexander F. Schier

Nature **411**, 607–610 (2001).

In Fig. 1 of this Letter, panels h, m and p were incorrectly labelled. The corrected panels are reproduced below. □



From cameras to Twisters

An eclectic haul — the latest from lab equipment manufacturers.

Gen II Multi Dimensional Imager

Universal Imaging www.image1.com

High-speed system images

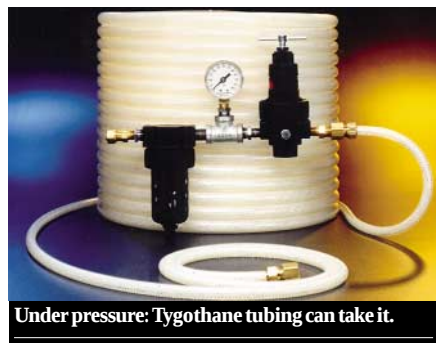
This micro imaging hardware platform, the Gen II Multi Dimensional Imager (MDI), was originally developed for high-throughput screening. It uses MetaMorph Imaging Software as the engine and is intended for virtually all multi-dimensional applications. These applications involve automated acquisition and review of digital microscopy images, including Z-axis or multiple focus series, multiple fluorochromes, time lapse and multiple stage positions. A high-speed dichroic changer and 10-position excitation and emission filter wheels operate simultaneously to enable rapid changes for fluorophore optimization and ultra-short exposure times. The focus motor moves in 30-ms steps for rapid Z-series data collection. Other MDI features include a 10-position neutral density filter wheel, a 5-position dichroic changer, 175 W xenon light source, a high speed focus device, automated 6-position objective turret and a motorized precision stage.

Amplification plates

Nalge Nunc International www.nalgenunc.com

Skirt the issue

NUNC Brand 96-well amplification plates for PCR fit a wide variety of thermal cyclers that hold an aluminium block for 96 V-shaped, 0.2 ml tubes. The plates are available with a full skirt or mini-skirt, complementing the original NUNC 96 amplification plate with no skirt. The plates can be sealed with 1 × 8 strip caps or amplification tape to limit evaporation and minimize contamination. Dome-shaped caps provide close contact with heated lids on thermal cyclers, remaining at a uniform temperature to eliminate condensation without requiring use of an oil overlay. Used in conjunction with the silicone spacer plate, the heat and pressure of the thermal cycler lid will mould the amplification tape around the wells, providing even heat transfer and a seal



Under pressure: Tygothane tubing can take it.



Platform change: Gen II from Universal Imaging, for high-throughput imaging.

against evaporation and well-to-well contamination. The plates are certified RNase and DNase-free and have 96 wells in an 8 × 2 × 12 array, each with a conical bottom and each holding a sample volume of 0.2 ml

Falcon imaging plates

BD Biosciences-Discovery Labware

www.bd.com

Through a glass, brightly

BD Falcon Glass-Bottom Imaging Plates are recommended for laser scanning and confocal microscopy applications. The use of high-quality glass ensures low autofluorescence while the overall flatness of each microplate ensures less error in scanning and imaging applications. The plates are available in 96-, 384- and 1,536-well formats.

Tygothane

Saint-Gobain Performance Plastics

www.sgpl.com

Tubing for harsh environments

Tygothane Precision Polyurethane Pressure Tubing, Formulation C-544-A IB, features inner braiding to provide the strength needed in high-pressure applications. The tubing is resistant to tearing and abrasion, oils, greases, and fuels. Tygothane retains flexibility in sub-zero environments. Typical applications include abrasive and viscous slurry transfer, lubrication and degreaser dispensing, pellet and powder transfer, coolant recovery systems, pneumatic sensory devices and instrumentation control lines.

Hybond Atlas

Amersham Pharmacia Biotech

www.apbiotech.com/hybondatlas

cDNA expression arrays for high throughput

Hybond Atlas cDNA expression arrays are aimed at high-throughput profiling of gene

expression. Produced for Amersham Pharmacia Biotech by Clontech Laboratories, each array contains 1,176 selected cDNA fragments arrayed on to nylon membranes, enabling the rapid assessment of the differential gene expression of hundreds of genes simultaneously. The range covers immobilized gene-specific cDNAs representing known genes that survey major biological pathways and functions and are available for human, mouse and rat research. The protocols have been specifically developed to work with Redivue ³²P or ³³P nucleotides for convenience in array applications and the kits include labelling reagents and ready-to-use hybridization buffer to increase sensitivity and reduce background. Hybond Atlas Arrays are designed to be accessible to all laboratories performing isotopic blot hybridization experiments, and are visualized using common techniques and routine lab equipment.

Wizard DNA system

Promega

www.promega.com

How does your garden grow?

The Wizard Magnetic 96 DNA Plant System automates DNA purification, for applications such as marker assisted breeding, seed quality analysis and GMO screening. MagneSil paramagnetic particles are used in a 96-well format in a system that can make use of a range of automation formats from a manual system using a multi-channel pipettor to a robotic system such as the Beckman Coulter Biomek 2000. Compatibility with 96-well automated grinding systems such as the SPEX CertiPrep Model 2000 Geno/Grinder provides additional automation and throughput. The chemistry of this system is optimized specifically for plant samples producing high-quality DNA from leaf and seed tissue. Promega states that validation of the system with corn and tomato leaf as well as canola and sunflower seed demonstrated reliable results with high reproducibility.

Chemical inventories

ChemSW

www.chemsw.com

Look, no wires

ChemSW now offers wireless palmtop scanning capability for chemical inventory management. The Symbol Technologies-based system uses radio frequency communication for immediate, real-time inventory updates whenever new chemicals are received or created. The Symbol Technologies wireless palmtop scanner is configured with the Palm Operating System and CIS Chemical Inventory Systems software from ChemSW. Operators can scan an arriving chemical into the inventory and label it immediately. The palmtop device has its own IP address, for a web-based link to the inventory database.

FLA-8000/FLA-5000

Fujifilm

www.fujimed.com

Two new scientific imaging systems

The Fujifilm FLA-8000 fluorescent image analyser combines fluorescent and radioisotope image detection in a single laser imaging system. It can capture images of samples labelled with fluorescent dyes or radioisotopes at high resolution. The FLA-8000 includes two standard fixed lasers: a second harmonic generation (SHG) green laser (532 nm) and a laser diode red laser (635 nm). An additional SHG laser (blue, 473 nm) is available as an option for additional applications. Typical imaging applications include: SYPRO Orange, SYPRO Ruby, SYBR Green I, FITC, Cy3, Rhodamine, SYPRO Red and CySTM. The new FLA-5000 imaging system allows scanning of fluorescent and radioisotopic samples as large as 40×46 cm at resolutions as low as $25 \mu\text{m}$. The system is especially suitable for protein electrophoresis in the proteomics field and general array analysis for genomics.

Spot Jr Slider Digital Camera

Diagnostic Instruments

www.diaginc.com

On the slide

Developed for microscopy applications, the Spot Jr Enhanced Slider digital camera utilizes the new Kodak KAF 040iE CCD. This full-frame CCD provides high sensitivity across the whole visible light spectrum, making it suitable for low light level and near IR imaging. A slider mechanism on the camera allows the colour filter to be moved out of the light path to maximize light throughput and increase sensitivity even further. Cooling to -37°C below ambient room temperature ensures low noise operation, allowing integration times up to 17 minutes. A fast focus window allows for easy isolation of regions of interest and critical focusing, and the click of a button snaps crisp, 12-bit images. Spot dedicated software is included with the camera, providing advanced editing tools, sequential imaging and measurement capabilities, calibration mark



A Spot of digital photography.

imprinting, database archiving, and report generation. The software works on both Mac and PC platforms and drivers for third party software packages are also available.

Cocktail selection utility

Packard Bioscience Company

www.packardinstrument.com

Cocktail time — and it's free

The FSA Cell and Cocktail Selection Guide is a software utility developed to make it easier to select the appropriate flow cell type and flow cocktail selection for flow scintillation analysis. This free utility is claimed to enable both beginners and advanced FSA users to select the most appropriate flow cell and flow cocktail for their needs. Additionally, the guide is capable of calculating optimal flow cell volumes. Also recently introduced is the Packard LSC Cocktail Guide, for cocktail selection for liquid scintillation counting. For a free copy of either of these two utilities on floppy disk, contact your local Packard office or download direct from the Packard web site.

syncroCOOL 435

Syncoscopy

www.syncroscopy.com

An all-in-one lab camera

Syncoscopy's syncroCOOL 435, "the only camera a lab will ever need", say the makers, combines the capture of high-resolution colour brightfield images with the same facilities for ultra-low noise, long integration imaging for fluorescence. The three $1,360 \times 1,024$ chips can output an image simultaneously, delivering a full resolution live image at 7.5 frames per second. This is suitable for routine brightfield slide scanning and focusing. Image capture is instantaneous. The camera's cooling mechanism takes the chip camera to at least

20°C below ambient, creating ultra-low noise images at all times. Integration periods can be well in excess of four minutes, making it possible to obtain bright images of even the faintest fluorescent stained samples in full colour.

KP-F102 camera

Hitachi Denshi

www.hitachi-denshi-uk.com

A square deal

The KP-F102 monochrome megapixel camera features partial scanning, and RS-232C remote control of the CPU to allow a wide range of parameters to be set remotely. The KP-F102 uses a CCD with $1,300 \times 1,030$ effective pixels and features frame on demand function. Pixel size is square, at $6.7 \times 6.7 \mu\text{m}$, and the makers claim a resolving power 3.5 times better than 400,000 pixel cameras. Possible applications include inspection, pattern recognition, high-resolution image filing, licence plate recognition and X-ray image processing and inspection.

Twister

Pro-Lab Diagnostics

www.pro-lab.com

A new angle on culture delivery

"Guaranteed authentic first generation cultures" is the aim of the Twister culture delivery system. Unlike swabs, plastic loops and discs, says makers Pro-Lab, the Twister guarantees that the culture is a "first generation". Cultures are from the NCIMB Ltd registered culture collection. Each Twister is made up of two chambers joined by a collar. One chamber contains a specially formulated "Resus Alpha" fluid, the second contains two microspheres coated with the bacterial or fungal strain of choice. When the Twister is activated by a simple twisting action, the culture is resuscitated ready for use. A wide range of cultures is available in the Twister format. The collar is colour-coded for ease of identification.

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENT

Polyclonal Antibodies & Custom Peptides

Please contact us about complete services for your custom peptide and polyclonal antibody needs.

ResGen™

Invitrogen Corporation

U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our web site at
www.resgen.com

Contacts

Publisher: Fabien Savenay

Editor: Paul Smaglik

Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building

4 Chancery Street

London N1 9XW, UK

Tel +44 (0) 20 7843 4961

Fax +44 (0) 20 7843 4996

e-mail: classified@nature.com

Group European Manager:

Leonie Welts (4954)

European Manager:

Nevin Bayoumi (4978)

European Sales Consultant:

Jonathan Bull (4973)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),

Andy Douglas (4975)

Holland/ Italy: Nevin Bayoumi (4978)

Scandinavia: Silje Opstrup (4994)

Production Manager:

Billie Franklin

To send films and materials use

London address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: classified@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau

Tel +33 (0) 1 43 20 16 51

Fax +33 (0) 1 43 20 51 52

e-mail: c.nioxchateau@nature.com

Spain/ Portugal:

Amanda Wren

Tel +34 91 391 3835

Fax +34 91 319 0091

e-mail: a.wren@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner

Tel +49 89 54 90 57 11/-2

Fax +49 89 54 90 57 20

e-mails: p.phelan@nature.com

k.turner@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: classified@nature.com

US Sales Director:

Ben Crowe

US Sales Manager:

Peyton Mason

US Advertising Coordinator:

Corissa Salzman

Japan Head Office, Tokyo

Shin-Mitsuke Building (4F),

3-6 Ichigaya Tamachi Shinjuku-Ku,

Tokyo 162

Tel +81 3 3267 8751

Fax +81 3 3267 8746

e-mail: k.cowan@nature.com

Japan Manager:

Kate Cowan

Stem-cell seduction

Roger Pedersen's decision to leave a faculty position at the University of California, San Francisco, for a similar post at Cambridge University because of Britain's more permissive policies towards stem-cell research turned more than a few heads last month (see *Nature* 412, 262; 2001). The United States prohibits federal funding of embryonic stem-cell derivation and research, whereas Britain allows both.

The move is especially noteworthy because of its timing — US President George W. Bush is set to announce whether or not he will reverse the country's policy. It is also noteworthy because other countries besides the United States — including Japan and Germany — are grappling with their own embryonic stem-cell policies. Embryonic stem-cell research is a political hot potato because deriving them from a young embryo effectively kills it.

The issue has been further muddled because the United States permits private funding of embryonic stem-cell research. But scientists such as Pedersen, who are funded by companies such as Geron, are unable to share their cell lines with publicly funded researchers. In the wake of Pedersen's move, Geron threatened to move some of its embryonic stem-cell research to Scotland. The company supports some of the United States' leading stem-cell researchers, including Jamie Thomson of the University of Wisconsin and John Gearhart of Johns Hopkins University — the two scientists who first published that they had isolated early forms of pluripotent stem cells in 1998.

Pedersen's defection sends a not-so-subtle signal: US scientists keen on following this promising avenue of research will leave the country for greener research pastures if the government persists. Geron's response emphasizes the point further by saying that companies, too, could take jobs to different countries where the research environment is more favourable.

Paul Smaglik

Naturejobs editor



Contents

MOVERS

AIDS virus co-discoverer leaves public research for private company; new head for Saturn mission; and more [Back page](#)

WWW.NATURE.COM/ NATUREJOBS

Career centre
Information on the scientific
job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS



BIOTECHNOLOGY

Jean-Claude Chermann, a member of the group that discovered the HIV virus, left INSERM, France's biomedical research agency, last month to head a lab for URRMA Biopharma in the south of France. Chermann decided to leave the government post because the new position would allow him more opportunities to transform his basic findings into products. "In France it's difficult when you are in public research to also work for a private company," Chermann says. His research over the past several years has focused on differences and commonalities between viruses in different people. "You have 14 million infected people, you have 14 million different viruses," he says. He discovered that some infected people don't progress to AIDS as quickly as others. He intends to develop two products based on that finding — a kit that will detect whether HIV-infected patients will quickly progress to AIDS, and antibodies against the virus for people resistant to the drug cocktails known as highly active anti-retroviral therapy. Before joining INSERM in 1988, Chermann spent 25 years at the Pasteur Institute, where he, with Luc Montagnier and others, discovered that a virus they then called LAV, caused AIDS.

PHYSICS

After several months as the head of a condensed matter and thermal physics group at Los Alamos National Laboratory, **Art Ramirez**, who joined Los Alamos from Bell Labs early this year, is in a good position to weigh the differences between public and privately funded science. On the one hand, at Bell, Ramirez never had to apply for funding or attend administrative meetings, as is now the case at the federal lab. On the other hand, his potential for funding and postdocs is much greater at Los Alamos, where he is focusing on material sciences — especially considering the financial woes of Lucent, which supports Bell Labs. And, with about 60 people, including 20 postdocs, in his lab, which he is still in the process of establishing, he has more technicians working with him. Ramirez jokes that the move ultimately resulted in one major trade-off: he now does less soldering, but more grant-writing.

US NATIONAL INSTITUTES OF HEALTH

Edmund Tramont, whose career included monitoring General Dwight D. Eisenhower's health and developing vaccines to prevent soldiers from contracting sexually transmitted diseases, became

director of the US National Institute of Allergies and Infectious Diseases (NIAID) Division of AIDS Research (DAIDS) last month. As director of DAIDS, Tramont will oversee an estimated \$444 million global research programme involving hundreds of clinical trials with the aim of treating, preventing and better understanding HIV/AIDS. After being drafted in 1968, Tramont spent 23 years in the army. He retired from it in 1991 and became professor at the University of Maryland Biotechnology Institute (UMBI), where he served as director of the institute's Medical Biotechnology Center. He continued his vaccine research and was involved in founding two biotech companies during that time. In 1998, Tramont became co-director of the Vaccine Division at the Institute of Human Virology, also part of the UMBI. Tramont replaces former DAIDS director John Killen, who has assumed the position of associate director for research ethics at NIAID.

ASTRONOMY

Last month **Carolyn Porco** was appointed an institute scientist in the Space Science and Engineering Division at Southwest Research Institute. Porco will lead the institute in the Cassini Saturn imaging mission, from institute facilities in Boulder, Colorado. The Cassini spacecraft, launched in 1997, will enter Saturn's orbit in 2004. Porco will direct remote sensing of the atmosphere, interior, rings, magnetosphere, its moon Titan and its icy moons. Before joining the institute, Porco taught and conducted research at the California Institute of Technology, the University of Arizona and the University of Colorado.

UNIVERSITY ADMINISTRATION

Four professors at Imperial College, London, took up newly created positions as 'principals of faculties' on August 1. **John Perkins**, former head of the Department of Chemical Engineering and Chemical Technology, became principal of the Faculty of Engineering; **Michael Hassell**, former head of the Department of Biology and Biochemistry, became principal of the Faculty of Life Sciences; **Leszek Borysiewicz**, principal of Imperial College School of Medicine, will retain that role and will also serve as principal of the Faculty of Medicine; and **John Pendry**, former head of the Department of Physics, became principal of the Faculty of Physical Sciences. The new positions were designed to encourage more interdisciplinary collaboration within the college.



Art Ramirez



Carolyn Porco



Edmund Tramont

CONTACTING US AT MOVERS

Please send any information on new appointments or job moves to: naturejobseditor@naturedc.com.